

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
27 December 2001 (27.12.2001)

PCT

(10) International Publication Number
WO 01/98454 A2

- (51) International Patent Classification⁷: C12N [DE/DE]; Grosse Lachstrasse 30a, 69207 Sandhausen (DE).
- (21) International Application Number: PCT/IB01/02050
- (22) International Filing Date: 25 April 2001 (25.04.2001)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data: 60/199,380 25 April 2000 (25.04.2000) US
- (63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application: US 60/199,380 (CIP) Filed on 25 April 2000 (25.04.2000)
- (81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.
- (84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- Published:
— without international search report and to be republished upon receipt of that report
- (71) Applicant (*for all designated States except US*): GERMAN HUMAN GENOME PROJECT [DE/DE]; Fraunhofer Patentstelle, Leonrodstrasse 68, 80636 Munich (DE).
- (72) Inventor; and
- (75) Inventor/Applicant (*for US only*): WIEMANN, Stefan
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: HUMAN DNA SEQUENCES

(57) Abstract: Novel human cDNA sequence of a clones, the encoded protein sequence of a clones, antibodies and variants thereof, are provided. The disclosed sequence of a clones find application in a number of ways, including use in profiling assays. In this regard, various assemblages of nucleic acids or proteins are provided that are useful in providing large arrays of human material for implementing large-scale screening strategies. The disclosed sequence of a clones may also be used in formulating medicaments, treating various disorders and in certain diagnostic applications.

Applicants: Paz Einat et al.
Serial No.: 10/618,408
Filed: July 11, 2003
Exhibit 5

WO 01/98454 A2

HUMAN DNA SEQUENCES

Background of the Invention

Current methods for testing pharmacological substances rely
5 on a three-stage testing approach to drug development. First,
candidate compounds are typically screened in some sort of *in*
vitro system, like inhibition of cancer cell growth. Candidates
are then tested in an animal model, as a first approximation of
systemic effects, including efficacy and toxicity. Compounds
10 that still show promise after these initial *in vivo* screens,
finally are tested in humans. Again, human testing typically
occurs in three phases: toxicity; preliminary efficacy; and
efficacy. The entire process can take more than a decade and
cost hundreds of millions of dollars. Aside from the monetary
15 costs and protracted time scale, moreover, current testing
regimes waste the lives of countless laboratory animals and
needlessly endanger the lives of human subjects.

A need exists, therefore, for more sophisticated drug
screening techniques that can be done rapidly *in vitro*. These
20 screening techniques ideally will be reflective of systemic
and/or organ-specific responses, so that they provide a reliable
indicator of action in a human body. Current techniques,
however, tend to utilize only a single or limited number of
markers, thus answering only very simple questions that are of
25 questionable medical import. For example, a typical *in vitro*
assay may ask whether a lead compound binds a particular
receptor, which has been implicated in a certain disorder. It is
presumed that such binding is indicative of therapeutic
usefulness, but it does not even purport to address systemic
30 effects.

Not only are screening techniques for efficacy inadequate,
the available toxicity screens likewise are inadequate.
Toxicity, on a first level, is usually measured by animal
testing. Aside from the complications related to *in vivo* versus
35 *in vitro* testing, such screens are insufficient because of
differences in metabolism, uptake, etc., relative to humans.

Thus, improved methods would be not only be *in vitro*-based, they would also be more "human."

With the increasing miniaturization of screening assays and the growing availability of targets for pharmaceutical intervention, there is increasing interest in developing arrays containing large numbers of these targets that can be assayed simultaneously. If such an array contains a large enough population of targets, it can be used to essentially mimic the systemic response. In other words, the array becomes an *in vitro* surrogate for the human body. The more refined the array, the more accurate the predictive capability. In theory, an array could be constructed that can detect all of the known human expression products simultaneously, thereby, providing a very reliable indicator of the human response to a given compound. These arrays offer advantages over the present *in vitro* screening systems in that they can assay large numbers of responses simultaneously. They are superior to animal testing because they are more "human" and, thus, more predictive of human responses.

In order to construct such arrays, however, the field is in need of further human targets. Advantageously, such targets will be provided with additional physiologically relevant information, such as whether the target is expressed in a particular tissue and whether it is related to a known functional class of targets. In this way, the artisan can focus as needed, for example, on tissue-specific effects or target class-specific effects, thereby providing information useful in evaluating efficacy and/or toxicity.

In addition to a need for pharmacological screening targets, there is a need for further pharmacological substances. These substances can be used in the formulation of medicinal compositions and in treating a wide variety of disorders.

The present invention responds to the aforementioned and other needs in the field by providing a population of novel targets useful, *inter alia*, in the profiling and medicinal contexts described above.

Summary of the Invention

It is an object of the invention, therefore, to provide a set of human cDNA clones. Further to this object, the invention provides sequences of human cDNA clones that were isolated from libraries generated from different human tissues.

5 It is another object of the invention to provide assemblages of targets useful in profiling matrices for screening pharmacological test compounds. According to this object, assemblages comprising different populations of human nucleic acids, proteins and antibodies are provided. In different
10 embodiments, cDNA library-specific assemblages and target-family-specific targets are provided.

It is a further object of the invention to provide a database of human nucleotide and protein sequences. Further to this object, novel human nucleotide and protein sequences are
15 provided in electronic form. In one embodiment, one or more of these sequences is provided in a searchable database.

It is still another object of the invention to provide biologically active target molecules useful in treating or detecting human disorders. Further to this object, the invention
20 provides nucleic acid and protein molecules that have the capacity to affect disease etiology or symptoms or correlate with known disease states. Also further to this object, a database is provided which comprises the disclosed molecules in electronic form.

25 Detailed Description

The invention results from a need in the art for new human nucleic acids and proteins. This need arises in several contexts. First, there is a need to identify targets for therapeutic intervention. Second, there is a need to identify molecules that
30 may be adversely affected in a therapeutic context, thereby resulting in toxicity. Knowledge of these molecules will aid in the design of new medicaments with enhanced efficacy and decreased toxicity. Finally, the need encompasses human nucleic acids and proteins that have medicinal applicability in their own right.

35 In view of these needs, the present inventors set out to isolate and sequence human cDNAs from tissue-specific libraries.

In this way, they represent subsets of molecules likely to be targets for therapeutic intervention or for avoiding toxicity. In addition, the inventors divided the molecules into various sub-categories, based on suspected functionality, structural
5 similarity etc, which are of interest from a pharmacological perspective.

GENERAL DESCRIPTION OF THE INVENTIVE MOLECULES

The present invention provides novel polynucleotide molecules that, in some instances, have similarities with known molecules.
10 The inventive DNAs were cloned from five different human cDNA libraries. In addition to these DNA molecules, the invention provides their protein translations and antibodies derived from them. The inventive DNA and protein sequences are shown individually in the Description of the Sequences. The inventive
15 nucleic acids also include the complements of the DNA sequences provided in the Description of the Sequences as well as their RNA counterparts. Methods of producing the molecules also are provided. Further, the invention provides methods for detecting all or part of the molecules and of detecting polynucleotides
20 encoding all or part of the molecules.

The inventive molecules derive from five cDNA libraries: human fetal brain; human fetal kidney; human melanoma; human testis; and human amygdala. For convenience, each sequence bears a designation that indicates from which library it is derived. In
25 particular, these designations are: "hfpbr" for human fetal brain; "hfkcd" for human fetal kidney; "hmel" for human melanoma; "htes" for human testis; and "hamy" for human amygdala. The individual libraries were constructed and screened as described below in the examples.

30 The protein and DNA molecules of the invention are variously described herein as "target" molecules or "inventive" molecules. The sequences and other information pertinent to the nucleic acid and protein molecules of the invention are shown below in the Description of the Sequences.

35

Description of the Sequences

Key to the Description of the Sequences

The descriptions below provide the coding sequences of the inventive cDNAs, as well as the protein sequences and other useful information, as set out herein.

5 Grouping

The clones were assigned to the following sixteen functional and/or tissue-derived groups:

1. Amygdala derived
2. Cell Cycle
3. Cell Structure and Motility
4. Differentiation/Development
5. Intracellular Transport and Trafficking
6. Melanoma derived
7. Metabolism
8. Nucleic Acid Management
9. Signal Transduction
10. Transmembrane Protein
11. Transcription Factors
12. Brain derived
13. Kidney derived
14. Mammary Carcinoma derived
15. Testes derived
16. Uterus derived

Description of Clone Files

The individual clone files are structured in the same pattern. The Sections are separated by paragraphs.

1. Clone Name

The clone names are deciphered with reference to the following example:

DKFZphfk2_3k1, wherein the code represents:

- producer of library ("DKFZ") (for convenience, this reference may be eliminated)
- a "p" for "plasmid cDNA library" (for convenience, this reference may be eliminated)
- library name (e.g. hfbr = human fetal brain; hfkd = human fetal kidney; hmel = human melanoma; htes = human testis; hamy = human amygdala)
- an underscore ("_") to separate library information from plate information
- plate number (e.g. "3")
- plate coordinates (letter first; e.g. "k12")

2. Group

3. Introduction

short review of the similarities, function of the protein and possible applications

4. Short Information

specifications about the cDNA (who sequenced, completeness of the cDNA, similarity, who sequenced, chromosomal localisation, length of cDNA, localisation of poly A tail and polyadenylation signal)

5. cDNA-Sequence

6. BLASTn Results

search results of blasting the cDNA sequence against all public databases

7. Medline Entries

information about genes/proteins similar to the novel cDNA (if available)

8. Putative Encoded Protein Information

specifications about the encoded protein (ORF: length and localisation of the reading frame)

9. Protein Sequence

10. BLASTp Results

search results of blasting the protein sequence against all public databases

11. Pedant Information

output of fully automated annotation: summarises peptide information, homologies, patterns as follows:

[[Length]]

- length of the protein = number of amino acid residues

[[MW]]

- molecular weight of the protein

[[pI]]

- isoelectric point

[[HOMOL]]

- shows protein with closest similarity to the cDNA-encoded protein

5 [[FUNCAT]]

- functional information according to a catalogue developed by Munich Information center for Protein Sequences (MIPS)

[[BLOCKS]]

10 - Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks for the Blocks Database are made automatically by looking for the most highly conserved regions in groups of proteins documented in the Prosite Database. The Prosite pattern for a protein group is not
15 used in any way to make the Blocks Database and the pattern may or may not be contained in one of the blocks representing a group. These blocks are then calibrated against the SWISS-PROT database to obtain a measure of the
20 chance distribution of matches. It is these calibrated blocks that make up the Blocks Database. The WWW versions of the Prosite and SWISS-PROT Databases that are used on this server are located at the ExPASy World Wide Web (WWW) Molecular Biology Server of the Geneva University Hospital
25 and the University of Geneva. World Wide Web URL http://blocks.fhcrc.org/blocks/about_blocks.html/ is the entry point to the database.

- here Blocks segments found in the analysed protein sequences are displayed

30 [[SCOP]]

Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The scop database provides a detailed and comprehensive description of the structural and
35 evolutionary relationships between all proteins whose structure is known, including all entries in Brookhaven National Laboratory's Protein Data Bank (PDB). It is available as a set of tightly linked hypertext documents which make the large database comprehensible and accessible.

In addition, the hypertext pages offer a panoply of representations of proteins, including links to PDB entries, sequences, references, images and interactive display systems. World Wide Web URL [http://scop.mrc-](http://scop.mrc-lmb.cam.ac.uk/scop/)

5. [lmb.cam.ac.uk/scop/](http://scop.mrc-lmb.cam.ac.uk/scop/) is the entry point to the database. Existing automatic sequence and structure comparison tools cannot identify all structural and evolutionary relationships between proteins. The scop classification of proteins has been constructed manually by visual inspection and comparison of structures, but with the assistance of
10 tools to make the task manageable and help provide generality. Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy, but the principal levels are family,
15 superfamily and fold. The exact position of boundaries between these levels are to some degree subjective. Scop evolutionary classification is generally conservative: where any doubt about relatedness exists, we made new divisions at the family and superfamily levels.

20 - - here SCOPE segments found in the analysed protein sequences are displayed

[[EC]]

ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the
25 recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided. World Wide Web URL <http://www.expasy.ch/enzyme/> is
30 the entry point to the database.

- here EC-number and name of enzymes with similarity to the analysed protein sequences are displayed

[[PIRKW]]

35 - functional information according to the Protein Information Resource (PIR) database catalogue developed by Munich Information Center for Protein Sequences (MIPS), the National Biomedical Research Foundation (NBRF) and the International Protein Information Database in Japan (JIPID).
[[SUPFAM]]

- information according to the Protein Information Resource (PIR) database catalogue of protein superfamilies developed by Munich Information Center for Protein Sequences (MIPS), the National Biomedical Research Foundation (NBRF) and the International Protein Information Database in Japan (JIPID).

[[PROSITE]]

please refer to 12. PROSITE Motifs

[[PFAM]]

please refer to 13. PFAM Motifs

[[KW]]

- overall 2dimensional folding information
- 3D indicates that the proteins is similar to a protein of which a 3 dimensional structure is known
- overall structural information

[[]]

The last PEDANT-block depicts information about the folding structure of the protein generated by PREDATOR. PREDATOR is a secondary structure prediction program. It takes as input a single protein sequence to be predicted and can optimally use a set of unaligned sequences as additional information to predict the query sequence. The mean prediction accuracy of PREDATOR is 68% for a single sequence and 75% for a set of related sequences. PREDATOR does not use multiple sequence alignment. Instead, it relies on careful pairwise local alignments of the sequences in the set with the query sequence to be predicted.

World Wide Web URL http://www.embl-heidelberg.de/argos/predator/predator_info.html is the entry point to the database.

- H = helix, E = extended or sheet, _ = coil, T = transmembrane, B = beta
- x indicates a low-complexity region with repeat-like structure which is omitted in all BLAST searches

12. PROSITE Motifs

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if

any) a new sequence belongs. World Wide Web URL
http://www.expasy.ch/prosite/ is the entry point to the database.
A description of the prosite consensus patterns is provided
herein, after the description of the individual sequences.

5

13. PFAM Motifs

PFAM (protein families) is a large collection of multiple
sequence alignments and hidden Markov models covering many common
protein domains. World Wide Web URL http://www.sanger.ac.uk/Pfam/
10 is the entry point to the database.

In the charts below, the groups of sequences are listed, and
the description of the individual clones follows.

Group Amygdala derived

CloneID	Homology	Function	Group
amy2_12g7	without similarity to known proteins	No informative BLAST results; No predictive prosite, pfam or SCOP motive.	amygdala derived
amy2_12i1	weak similarity to F41E6.3 of <i>Caenorhabditis elegans</i>	No informative BLAST results; No predictive prosite, pfam or SCOP motive	amygdala derived
amy2_13g19	without similarity to known proteins	The novel protein contains a PROSITE ASP_PROTEASE motif and seem to be expressed ubiquitously.	amygdala derived
amy2_16e14	similar to carbonic anhydrase-related proteins	No informative BLAST results; No predictive prosite, pfam or SCOP motive. A similar cDNA encoding a protein of the same length was identified in sheep. This protein shows a strong signal sequence, which indicates that it is a secreted protein. The new protein belongs to a protein family, which was designated carbonic anhydrase-related protein XI (CA-RP XI), encoded by CA11 (human) and CA11 (mouse, rat). Despite potentially inactivating changes in the active-site residues, CA-RP XI is evolving very slowly in mammals, a property indicative of an important function, which has also been observed in the two other "acatalytic" CA isoforms, CA-RP VII and CA-RP X. No informative BLAST results; No predictive prosite, pfam or SCOP motive	amygdala derived
amy2_24k15	weak similarity to pecanex of <i>Drosophila melanogaster</i> .	Pecanex is a maternal-effect neurogenic gene, involved in differentiation processes in the developing central nervous system. DKFZpham2_24k15 seems to be expressed ubiquitously.	Amygdala derived
amy2_2a13	without similarity to known proteins	No informative BLAST results; No predictive prosite, pfam or SCOP motive	amygdala derived
amy2_2i17	without similarity to known proteins	No informative BLAST results; No predictive prosite, pfam or SCOP motive. Most ESTs are derived from brain and pancreas No informative BLAST results; No predictive prosite, pfam or SCOP motive.	amygdala derived

Group Brain derived

CloneID	Homology	Function	Group
DKF2ph...			
fbr2_7dd16	weak similarity to a human putative mitogen-activated protein kinase kinase kinase	No informative BLAST results; No predictive prosite, pfam or SCOP motife.	brain derived
fbr2_7de16	without similarity to known proteins.	The mRNA is differentially polyadenylated. No informative BLAST results; No predictive prosite, pfam or SCOP motife.	brain derived

Group cell cycle

CloneID	Homology	Function	Group
any2_121m2	Similarity to human PA26-T2 protein.	PA26-T2 is a p53 responsive gene. The protein is predominantly expressed in brain, breast and kidney and may represent a potential novel regulator of cellular growth. Isoforms are differentially induced by genotoxic stress (UV, gamma-irradiation and cytotoxic drugs) in a p53-dependent manner.	cell cycle
any2_2404	Similarity to human STIM1	The stromal interaction molecular 1 gene (STIM1) encodes a type I trans-membrane protein of unknown function, which induces growth arrest and degeneration of the human tumor cell lines G401 and RD but not HBL100 and Calu-6, suggesting a role in the pathogenesis of rhabdomyosarcomas and rhabdoid tumors. There is also strong similarity to a Mus musculus stromal cell protein, which selectively increases interleukin 7-dependent proliferation of pre-B cells. The novel protein contains 1 transmembrane domain.	cell cycle

Group Cell structure and motility

CloneID	Homology	Function	Group
DKFZph... amy2_121119	high similarity to a Rat ankyrin binding glycoprotein-1 related mRNA.	Ankyrin binding glycoproteins play a role in neural cell adhesion and in prostate tumor cell transformation. DKFZphamy2_121119 is expressed in brain, uterus and prostate above average	cell structure and motility
tes3_1bb5	similarity to various tropomyosins.	Tropomyosins play regulatory roles in cellular structure and transport.	cell structure and motility

Group Differentiation/Development

CloneID	Homology	Function	Group
DKFZph... amy2_1i24	partial similarity to rattus norvegicus Notch2 protein	Notch family molecules are thought to be negative regulators of neuronal differentiation in early brain development. Notch2 is expressed not only by neuronal cells in the embryonic brain, but also by glial cells in the postnatal brain.	differentiat ion/developm ent
amy2_1j14	high similarity to the allograft inflammatory factor-1 of Cyprinus carpio.	Allograft inflammatory factor-1 (AIF-1) is a protein involved in allograft rejection. In experimental autoimmune encephalomyelitis (EAE), neuritis(EAN) and uveitis (EAU) it is produced by macrophages and microglia cells.	differentiat ion/developm ent
amy2_2b14	Originates from TXBP151 mRNA by alternative splicing	It is ubiquitously expressed. The mRNA is also subject to alternative polyadenylation. Overexpression of TXBP151 in NIH3T3 cells causes inhibition of apoptosis induced by tumour necrosis factor (TNF). It binds to A20, which is also an inhibitor of cell death by a yet unknown mechanism.	differentiat ion/developm ent
amy2_7j5	similarity to Tspyl1 testis-specific Y-encoded-like protein of Mus musculus	TSPY genes are arranged in clusters on the Y chromosome of many mammalian species. TSPY is believed to function in early spermatogenesis and is a candidate for GSY, the putative gonadoblastoma-inducing gene on the Y. The TSPY family forms part of a superfamily, TISN, with autosomal representatives, highly conserved in mammals and beyond.	differentiat ion/developm ent

Group Intracellular Transport and Trafficking

CloneID	Homology	Function	Group
amy2_14b5	Shows 61% identity to the human TYL protein and 48% identity to the human Tic protein	Both proteins show similarity to Sac7 of <i>Saccharomyces cerevisiae</i> , which takes function in vesicular trafficking. The new protein shows also significant similarity to human ARN03, which is involved in the control of Golgi structure and function. JKFZphamy2_14b5 is predominantly expressed in the CNS and germ cells.	Intracellular r transport and trafficking
amy2_20b3	high similarity to murine synaptotagmin 3.	The novel protein contains two C2 domains. The C2 domain is thought to be involved in calcium-dependent phospholipid binding. Synaptotagmins are essential for Ca(2+)-regulated exocytosis of neurosecretory vesicles	Intracellular r transport and trafficking
fkcd2_3k1	very similar to rat testicular dynamin	Dynamin is a microtubule-associated force-producing protein, which is involved in the production of microtubule bundles and which is able to bind and hydrolyze GTP and provides the motor for vesicular transport during endocytosis. The protein is ubiquitously expressed, but in brain and testis above average.	Intracellular r transport and trafficking
mel2_7g14	Similarity to the dor (deep orange) protein of <i>Drosophila melanogaster</i> .	The novel protein is also similar to the vakuolar membrane protein pep3 of <i>Saccharomyces cerevisiae</i> , which is involved in protein sorting mechanisms. The expression profile is ubiquitous and a role in protein transport/targeting is likely.	Intracellular r transport and trafficking

Group Melanoma derived

CloneID	Homology	Function	Group
DKFZpg...			
mel2_12j1	similarity to integrin I of Saccharomyces cerevisiae	The novel protein contains a leucin zipper. No informative BLAST results; No predictive prosite, pfam or SCOP motive.	melanoma derived
mel2_7k19	without similarity to known proteins	Transcripts can be found in almost any tissue, but are most abundant in kidney and retina. No informative BLAST results; No predictive prosite, pfam or SCOP motive.	melanoma derived

Group Metabolism

CloneID DXYZphr.1:	Homology	Function	Group
amy2_2c22	similarity to the 1-acyl-glycerol-3-phosphate acyltransferase of Zea mays.	It contains one leucine zipper. The protein is believed to play a role in fatty acid metabolism. It is ubiquitous expressed, with a slight predominance in uterus, placenta and foreskin.	metabolism
fbr2_7a121	Similarity to beta-aspartate methyltransferases.	The L-isoaspartyl methyltransferase (Pimt), as an example, is a highly conserved enzyme utilising S-adenosylmethionine (AdoMet) to methylate aspartate residues of proteins damaged by age-related isomerisation and deamidation.	metabolism

Group Nucleic acid management

CloneID	Homology	Function	Group
amy2_1ln4 DKF2ph...	similarity to RAD16 of Schizosaccharomyces pombe and YLR363w of Saccharomyces cerevisiae.	The novel protein contains a ATP/GTP-binding site motif A (P-loop). It has similarity to RAD16 acts in a DNA repair pathway for removal of UV-induced DNA damage. YLR363w of Saccharomyces cerevisiae is a recombination repair protein	nucleic acid management
amy2_1l1	similarity to the murine hemin-sensitive initiation factor 2.	The hemin-sensitive initiation factor 2 is expressed predominantly in liver, spleen, colon and uterus and contains 2 protein kinase motifs. The mouse homologue inhibits protein synthesis in stress conditions by phosphorylation of eIF-2-alpha.	nucleic acid management
amy2_2gl2	similarity to NVL-2 of Rattus norvegicus.	The novel protein contains 3 EF-hand calcium-binding domains. The related human VILIP Ca-dependent protein specifically binds the 3'-untranslated region of the neurotrophin receptor, trkB, an mRNA localized to hippocampal dendrites in an activity-dependent manner. The new protein exhibits elevated expression in brain and testis.	nucleic acid management
fbr2_7ac12	high similarity to glutamyl-tRNA (Gln) amidotransferase subunit A of the hyperthermophilic bacterium Aquifex aeolicus.	The novel protein contains one ATP/GTP-binding site motif A (P-loop). This loop interacts with one of the phosphate groups of a A or G nucleotide. It is found in numerous ATP- or GTP-binding proteins, such as ATP synthase alpha and beta subunits, Myosin heavy chains, Kinesin heavy chains and kinesin-like proteins, dynamins and dynamin-like proteins, several kinases, DNA and RNA helicases, GTP-binding elongation factors and the Ras family of GTP-binding proteins. The protein seems to be expressed ubiquitously.	nucleic acid management
tes3_10i16	similarity to human ZK1.	The ZK1 gene is one of early response genes by exposure to ionizing radiation, and plays a role in radiation-induced apoptotic cell death on hematopoietic cells. The novel protein contains 16 zinc finger domains, a RGD cell attachment and a ATP GTP A domain.	nucleic acid management
tes3_3la10	similarity to histone H1 of Drosophila hydei.	Histone H1 variants are known to act as specific regulators of genes via the differential condensation of DNA.	nucleic acid management

Group Signal transduction

CloneID DEF2p...	Homology	Function	Group
amy2_10h17	weak similarity to murine hac1	The novel protein contains a Zinc finger motif of the C3HC4 type (RING finger). The RING-finger domain is involved in mediating protein-protein interactions. Proteins containing a RING-finger are: mammalian V(D)J recombination activating protein (RAG1), mouse rpt-1, human rfp, human 52 Kd Ro/SS-A protein and others. The family of RING finger proteins contains a number of oncogenes. For example PML, a probable transcription factor, BRCA1, the mammalian cbl- and bml-1 proto-oncogenes.	signal transduction
amy2_10p7	similarity to Na ⁺ /Ca ²⁺ exchange proteins	The Transport of Ca ²⁺ from the sarcoplasm into the sarcoplasmic reticulum is an essential process in the initiation of muscle relaxation. In addition, the novel protein contains a PROSITE multicopper oxidase signature. Multicopper oxidases are enzymes that possess three spectroscopically different copper centers.	signal transduction
amy2_12d7	a so far unknown alternative spliced form of disks large homolog DLG2.	It seems to be predominantly expressed in the retina, germ cells and brain. It contains a SH3-domain and a guanylate kinase domain. These conserved regions are shared among members of the discs-large family of proteins that include human p55, a membrane protein expressed in erythrocytes, rat PSD-95/SAP90, a synapse protein expressed in brain, Drosophila dig-A, a septate junction protein expressed in various epithelia, and human and mouse ZO-1 and canine ZO-2, two tight junction proteins. The Homologue of Drosophila dig-A, acts as a tumor suppressor. All members of this family may be involved in signal transduction.	signal transduction
amy2_2f1d6	Similarity to sodium channel protein beta1 of Rattus norvegicus.	The sodium channel protein beta 1 of Rattus norvegicus is crucial in the assembly, expression, and functional modulation of the heterotrimeric complex of the rat brain sodium channel. The expression of the new protein seems to be restricted to brain, all matching ESTs isolated so far, derive from there.	signal transduction
tes3_11c22	Partial similarity to mouse PC22b	The novel protein contains WD-repeats. WD-repeat proteins are known as regulatory elements in a large variety of pathways. The repeats form a propeller like structure, which serves as a platform for protein/protein interaction. The new protein is ubiquitously expressed, indicating that it takes an essential regulatory function in the cell.	signal transduction
tes3_11d21	Contains the full coding sequence of the human Nedd-4-like ubiquitin-protein ligase.	The novel protein contains four WW domains. The WW/rsp5/UBP domain has been shown to bind proteins with particular proline-motifs, and thus resembles somewhat SH3 domains. It is frequently associated with other domains typical for proteins in signal transduction processes. There is also a ubiquitin-protein ligase activity reported. The protein is believed to play an important role in protein-degradation pathways.	signal transduction
tes3_29f24	Similarity to murine net1a.	The closely related mNET1 activates signalling pathways in addition to those directly controlled by activated RhoA. The novel protein is expressed ubiquitously.	signal transduction
tes3_31j20	contains a Protein phosphatase 2C motif.	The novel protein shares 95% identity with the rat protein phosphatase 2C and is expressed ubiquitously. PP2C is a structurally diversified protein phosphatase family with a wide range of functions in cellular signal transduction. The transcription of the PP2Cdelta gene was activated in response to stress, like alcohol or UV irradiation. PP2C plays a role in cell cycle control.	signal transduction
tes3_5k22	similarity to human paraneoplastic neuronal antigen MAL	Antibodies against MAL were found in patients with paraneoplastic neurological disorders. The protein is predominantly expressed in testis and brain, but ESTs are also found in liver, lung uterus and kidney	signal transduction

Group Testis derived

CloneID DKZPg...	Homology	Function	Group
Tes3_10n10	without similarity to known proteins.	The mRNA is differentially polyadenylated and the novel protein is ubiquitously expressed. No Informative BLAST results; No predictive prosite, pfam or SCOP motive. No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
Tes3_11e17	without similarity to known proteins.		testis derived
Tes3_13d16	without similarity to known proteins	The EST-distribution signifies an ubiquitous expression pattern. No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
Tes3_14l17	without similarity to known proteins.	The mRNA is transcribed ubiquitously. No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
Tes3_15n14	weak similarity to the neurofilament triplet M protein of the rat.	Neurofilaments are the intermediate filaments specific to nervous tissue. They are probably essential to the tensile strength of the neuron, as well as to transport of molecules and organelles within the axon. Until now, ESTs of the novel mRNA could only be isolated from testes, germ cells and uterus. No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
Tes3_16p3	without similarity to known proteins	The novel protein is glutamine rich and contains a cell attachment RGD motif. According to the low number of ESTs and their origin the protein seems to be expressed ubiquitously at low levels. No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
Tes3_14p12	without similarity to known proteins	No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
Tes3_24k14	without similarity to known proteins	No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
Tes3_22i11	weak similarity to RCC1-like G exchanging factor RL6, UVR6 (UVB-resistance protein) of Arabidopsis thaliana and to the murine retinitis pigmentosa GTPase regulator.	No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
Tes3_22l24	Similarity to the F-box protein FBL2 of the rat.	No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
tes3_24g3	Without similarity to known proteins.	No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived
tes3_30p6	without similarity to known proteins.	No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	testis derived

Group Transmembrane proteins

CloneID DKZPg...	Homology	Function	Group
amy2_11d2	Without similarity to known proteins	The novel protein contains 2 transmembrane regions. No Informative Blast results; No predictive prosite, pfam or scope motive.	transmembran e proteins
amy2_12l017	without similarity to known proteins.	The novel protein contains 1 transmembrane region. No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	Transmembran e proteins
amy2_11l4	Similarity to the human 1(3)mbt protein homolog.	Mutations of the Drosophila 1(3)mbt gene lead to malignant brain tumors. The protein contains one transmembrane domain No Informative BLAST results; No predictive prosite, pfam or SCOP motive	Transmembran e proteins
amy2_24c6	without similarity to known proteins	The novel protein contains 1 transmembrane region. No Informative BLAST results; No predictive prosite, pfam or SCOP motive.	Transmembran e proteins

rbr2_76d4	without similarity to known proteins.	The novel protein contains 1 transmembrane region and a Cytochrome c family heme-binding site. No informative BLAST results; No predictive prosite, pfam or SCOP motive.	Transmembrane proteins
tes3_33a37	without similarity to known proteins	The novel protein contains 2 transmembrane regions and one leucine zipper. The protein is ubiquitously expressed with higher abundance in stomach, brain and testis. No informative BLAST results; No predictive prosite, pfam or SCOP motive.	Transmembrane proteins
tes3_17i23	without similarity to known proteins	The novel protein contains 2 transmembrane regions. ESTs can be found in testis, retina and brain. No informative BLAST results; No predictive prosite, pfam or SCOP motive.	Transmembrane proteins
tes3_20h12	without similarity to known proteins	The novel protein contains 1 transmembrane region and two leucine zippers. No informative BLAST results; No predictive prosite, pfam or SCOP motive.	Transmembrane proteins
tes3_7n32	without similarity to known proteins	The novel protein contains 1 transmembrane domain. No informative BLAST results; No predictive prosite, pfam or SCOP motive.	Transmembrane proteins
tes3_9e16	without similarity to known proteins	The novel protein contains 1 transmembrane region. The only EST described so far is from testis. No informative BLAST results; No predictive prosite, pfam or SCOP motive.	Transmembrane proteins

Group Transcription factors

CloneID	Homology	Function	Group
any2_34m1b	similarity to the homeotic protein emx2 of man, mouse and zebra fish as well as to the gene "empty spiracles" of <i>Drosophila melanogaster</i> .	Homeobox genes are known to play important roles in developmental processes. In zebrafish <i>emx2</i> mRNAs are found in the dorsal telencephalon, parts of the diencephalon and the otocyst. The human homologue <i>Emx2</i> appears to be already expressed in 8.5 day embryos. It is also expressed in the presumptive cerebral cortex, olfactory bulbs, in some neuroectodermal areas in the embryonic head including olfactory placodes in earlier stages and olfactory epithelia later in development. Mutants of the <i>D. melanogaster</i> gene "empty spiracles" display spiracles devoid of flizkorper, no antenna and an open head.	transcription factors
any2_1cl2	partial identity to I-kappa-B-related protein and to BRCA1.	I-kappa-B-related protein interacts with transcription factors and BRCA1 has a function in DNA damage response. I-kappa-B-alpha mutations contribute to constitutive NF-kappaB activity in cultured and primary HRS (Hodgkin/Reed-Sternberg) cells and are therefore involved in the pathogenesis of Hodgkin's disease (HD) patients	transcription factors
any2_2f22	similarity to YDL153c of <i>Saccharomyces cerevisiae</i>	The novel protein is ubiquitously expressed. YDL153C is involved in transcriptional silencing.	transcription factors
tes3_16n14	similarity to human giantin.	Giantin is discussed as an autoantigen in rheumatoid arthritis. The novel protein contains a leucine zipper and a putative Helix-loop-helix DNA-binding domain. Therefore it might be a novel transcription factor. Most EST hits are from testis and germ cells.	Transcription factors

DKFZphamy2_10h17

group: signal transduction

10 DKFZphamy2_10h17 encodes a novel 180 amino acid protein which shows weak similarity to murine hac1.

15 The novel protein contains a Zinc finger motif of the C3HC4 type (RING finger). The RING-finger domain is involved in mediating protein-protein interactions. Proteins containing a RING-finger are: mammalian V(D)J recombination activating protein (RAG1), mouse rpt-1, human rfp, human 52 Kd Ro/SS-A protein and others. The family of RING finger proteins contains a number of
 20 oncogenes. For example PML, a probable transcription factor, BRCA1, the mammalian cbl- and bmi-1 proto-oncogenes.

25 The new protein can find application in modulating protein-protein-interaction and in studying the expression profile of amygdala-specific genes.

weak similarity to hac1 (Mus musculus)

Sequenced by LMU

30

Locus: unknown

Insert length: 835 bp

Poly A stretch at pos. 751, polyadenylation signal at pos. 729

35

```

      1 CACAGAGATC ATTGTCAACC AGGCCTGTGG GGGGGACATG CCTGCCTTGG
    51 AAGGGGGACC CCATACCCCG CCACTGCCAC GCGGGCCCCG TAAGGGAAGC
   101 TCGGAGCTGG GCTTTCCCCG CGTGGCCCCA GAGGATGAGG TCATTGTGAA
   40 151 TCAGTACGTG ATTCGGCCTG GCCCCTCGGC CTCGGCGGCT TCTTCGGCGG
   201 CGGCAGGCGA GCCCCGCGT TGCCCCACCT GTGGGCACTC CTACAATGTC
   251 ACCCAGCGGA GGGCCCGCGT GCTGTCTGCT CTGCACTCTG TGTGTGAGCA
   301 GTGCCTGCAG ATTCTCTACG AGTCCTGCCC CAAGTACAAG TTCATCTCCT
   351 GCCCCACCTG CCGCCGTGAG ACTGTGCTCT TCACCGACTA CGGCCTGGCC
   45 401 GCGCTGGCTG TCAACACGTC CATCCTGAGC CGCCTGCCGC CTGAGGCGCT
   451 GACGGCCCCA TCCGGGGGTC AGTGGGGGGC TGAGCCCGAG GGCAGCTGCT
   501 ACCAGACCTT CCGGCAGTAC TGTGGGGCCG CGTGACCTG CCACGTGCGG
   551 AACCCTACTG CCGCCTGCTC CATCATGTAG TAGCGCCTGC CTGCCCGCCA
   601 CTGCCCGCTG AGCCTCGCTC GCTGCTTCTT CAGGGACCCG GCCCTGCCCT
   50 651 GCGCCCCGCT GACCCTTCCT TCCCCACCAT GGCTTCCGGC CCCACCCCGA
   701 GTGGCATTGT CGCTGCAGCC AACTTTGCCA TTAATACTCT TTGCCAAAGT
   751 TAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
   801 AAAAAAAAAA AAAAGAAAAA AAAAAAAAAA AAAAG

```

55

BLAST Results

No BLAST result

Medline entries

5

No Medline entry

10

Peptide information for frame 2

ORF from 38 bp to 577 bp; peptide length: 180
 Category: similarity to unknown protein
 Classification: Cellular transport and traffic
 Prosite motifs: PRENYLATION (177-180)
 ZINC_FINGER_C3HC4 (81-90)

20

1 MPALEGAPHT PPLPRRPRKG SSELGFPRVA PEDEVIVNQY VIRPGPSASA
 51 ASSAAAGEPL ECPTCGHSYN VTQRRPRVLS CLHSVCEQCL QILYESCPKY
 101 KFISCPTCRR ETVLFTDYGL AALAVNTSIL SRLPPEALTA PSGGQWGAEP
 151 EGSCYQTFRQ YCGAACTCHV RNPLSACSIM

25

BLASTP hits

30 No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_10h17, frame 2

No Alert BLASTP hits found

35

Pedant information for DKFZphamy2_10h17, frame 2

Report for DKFZphamy2_10h17.2

40

[LENGTH] 180
 [MW] 19400.27
 [pI] 7.95
 45 [HOMOL] TREMBL:AC007727_7 gene: "F8K7.7"; Arabidopsis
 thaliana chromosome 1 BAC F8K7 sequence, complete sequence. 3e-06

50

[BLOCKS] BL00839C
 [BLOCKS] PF01462A
 [BLOCKS] PR00763H
 [BLOCKS] BL00518 Zinc finger, C3HC4 type, proteins
 [PROSITE] PRENYLATION 1
 [PROSITE] ZINC_FINGER_C3HC4 1
 [PFAM] Zinc finger, C3HC4 type (RING finger)
 55 [KW] Alpha_Beta
 [KW] LOW_COMPLEXITY 5.56 %

SEQ MPALGAPHTPPLPRRPRKGSSELGFPRVAPEDDEVIVNQYVIRPGPSASAASSAAAGEPL
 SEGxxxxxxxxxxxx.....
 PRD ccc

5 SEQ ECPTCGHSYNVTQRRPRVLSCLHSVCEQCLQILYESCPKYKFISCPTCRRETVLFTDYGL
 SEG
 PRD ccc

10 SEQ AALAVNTSILSRLPPEALTAPSGGQWGAPEEGSCYQTFRQYCGAACTCHVRNPLSACSIM
 SEG
 PRD cchhhhhhhhhcc

15 Prosite for DKFZphamy2_10h17.2

PS00294	177->181	PRENYLATION	PD0C00266
PS00518	81->91	ZINC_FINGER_C3HC4	PD0C00449

20

Pfam for DKFZphamy2_10h17.2

25 HMM_NAME Zinc finger, C3HC4 type (RING finger)

HMM
 *CPICFCTFQ1DyPWPfdePmM1PCgHsFCypCIrrW.....C
 CP C Y+ +P+ L C+HS C+ C+ ++

30 C
 Query 62 CPTC----GHSYNVTQRRPRVLSCLHSVCEQCL-
 QILYESCPKYKFISC 105

35 HMM PmC*
 P C
 Query 106 PTC 108

5 group: signal transduction

DKFZphamy2_10p7 encodes a novel 1615 amino acid protein with similarity to Na⁺/Ca²⁺ exchange proteins.

10 The Transport of Ca²⁺ from the sarcoplasm into the sarcoplasmic reticulum is an essential process in the initiation of muscle relaxation.

15 In addition, the novel protein contains a PROSITE multicopper oxidase signature. Multicopper oxidases are enzymes that possess three spectroscopically different copper centers.

The new protein can find application in modulation of Na⁺/Ca²⁺-exchange and voltage-dependend processes.

20 similarity to Na⁺/Ca²⁺ exchange proteins

ATG in frame 3 is first in clone.

25 Sequenced by LMU

Locus: unknown

Insert length: 5236 bp

30 Poly A stretch at pos. 5216, no polyadenylation signal found

```

      1 CGGACGCGTG GCGGACGCG TGGGCCCTGT ATACCTGTGC CACTTTGTGC
      51 CTTAAGGAAC AAGCTTGCTC AGCGTTTTCA TTTTTCAGTG CTCTGAGGG
35    101 TCCCCAGTGT TTCTGGATGA CATCATGGAT CAGCCCAGCT GTCAACAATT
      151 CAGACTTCTG GACCTACAGG AAAAACATGA CCAGGGTAGC ATCTCTTTTT
      201 AGTGGTCAGG CTGTGGCTGG GAGTGACTAT GAGCCTGTGA CAAGGCAATG
      251 GGCCATAATG CAGGAAGGTG ATGAATTGCG AAATCTCACA GTGTCTATTC
      301 TTCCTGATGA TTTCCAGAG ATGGATGAGA GTTTTCTAAT TTCTCTCCTT
40    351 GAAGTTCACC TCATGAACAT TTCAGCCAGT TTGAAAAATC AGCCAACCAT
      401 AGGACAGCCA AATATTTCTA CAGTTGTGAT AGCACTAAAT GGTGATGCCT
      451 TTGGAGTGTT TGTGATCTAC AGTATTAGTC CCAATACTTC CGAAGATGGC
      501 TTATTTGTTG AAGTTCAGGA GCAGCCCCAA ACCTTGGTGG AGCTGATGAT
      551 ACACAGGACA GGGGGCAGCT TAGGTCAAGT GGCAGTCGAA TGGCGTGTTG
45    601 TTGGTGGAAC AGCTACTGAA GGTTTAGATT TTATAGGTGC TGGAGAGATT
      651 CTGACCTTTG CTGAAGGTGA AACCAGAAAG ACAGTCATTT TAACCATCTT
      701 GGATGACTCT GAACCAGAGG ATGACGAAAG TATCATAGTT AGTTTGGTGT
      751 ACACTGAAGG TGGAAGTAGA ATTTTGCCAA GCTCCGACAC TGTTAGAGTG
      801 AACATTTTGG CCAATGACAA TGTGGCAGGA ATTGTTAGCT TTCAGACAGC
50    851 TTCCAGATCT GTCATAGGTC ATGAAGGAGA AATTTTACAA TTCCATGTGA
      901 TAAGAACTTT CCCTGGTCGA GGAAATGTTA CTGTAACTG GAAAATTATT
      951 GGGCAAAATC TAGAACTCAA TTTTGCTAAC TTTAGCGGAC AACTTTTCTT
     1001 TCCTGAGGGG TCGTTGAATA CAACATTGTT TGTGCATTTG TTGGATGACA
     1051 ACATTCCTGA GGAGAAAGAA GTATACCAAG TCATTCTGTA TTGTGTCAGG
55    1101 ACACAAGGAG TTCCACCAGC CGGAATCGCC CTGCTTGATG CTCAAGGATA
     1151 TGCAGCTGTC CTCACAGTAG AAGCCAGTGA TGAACCACAT GGAGTTTTAA
     1201 ATTTTGCTCT TTCATCAAGA TTTGTGTTAC TACAAGAGGC TAACATAACA
     1251 ATTCAGCTTT TCATCAACAG AGAATTTGGA TCTCTAGGAG CTATCAATGT
```

1301 CACATATACC ACGGTTCTCTG GAATGCTGAG TCTGAAGAAC CAAACAGTAG
1351 GAAACCTAGC AGAGCCAGAA GTTGATTTTG TCCCTATCAT TGGCTTTCTG
1401 ATTTTAGAAG AAGGGGAAAC AGCAGCAGCC ATCAACATTA CCATTCTTGA
1451 GGATGATGTA CCAGAGCTAG AAGAATATTT CCTGGTGAAT TTAACCTACG
5 1501 TTGGACTTAC CATGGCTGCT TCAACTTCAT TTCCTCCCAG ACTAGATTCA
1551 GAAGGTTTGA CTGCACAAGT TATTATTGAT GCCAATGATG GGGCCCGAGG
1601 TGTAATTGAA TGGCAACAAA GCAGGTTTGA AGTAAATGAA ACCCATGGAA
1651 GTTTAACATT GGTAGCCCAG AGGAGCAGAG AACCTCTTGG CCATGTTTCC
1701 TTATTTGTGT ATGCTCAGAA TTTGGAAGCA CAAGTGGGGC TGGATTATAT
10 1751 CTTACCCCCA ATGATTCTTC ATTTTGCTGA TGGAGAAAGG TATAAAAATG
1801 TCAATATCAT GATTCTTGAT GATGACATTC CAGAAGGAGA TGAAAAATTT
1851 CAGCTGATTT TAACAAATCC TTCTCCTGGA CTAGAGCTAG GGAAAAATAC
1901 AAGTGCCTTA ATTATTGTCT TTGCTAATGA TGACGGCCCT GGAGTTCTAT
1951 CATTTAACAA CAGTGAGCAC TTTTTCCTAA GAGAGCCAAC AGCTCTCTAC
15 2001 GTCCAGGAGA GTGTTGCAGT ATTGTACATT GTTCGGGAAC CTGCACAAGG
2051 ATTGTTTGGG ACAGTGACAG TTCAGTTCAT TGTGACAGAA GTGAATTCCT
2101 CAAATGAATC TAAAGATCTG ACTCCTTCCA AAGGCTATAT TGTTTTAGAA
2151 GAAGGTGTTT GATTCAAGGC CCTACAAATA TCTGCCATAT TAGACACGGA
2201 ACCAGAAATG GATGAGTATT TTGTTTGCAC CTTGTTTAAAT CCAACTGGAG
20 2251 GTGCTAGACT AGGGGTGCAT GTTCAAACCC TGATAACAGT TTTGCAAAAC
2301 CAGGCCCCCT TGGGGCTATT CAGTATCTCT GCAGTTGAAA ATAGAGCCAC
2351 CTCCATAGAC ATCGAAGAAG CCAATAGGAC CGTGTATTTA AATGTATCTC
2401 GAACTAATGG CATTGATTTG GCTGTGAGTG TGCAGTGGGA GACAGTATCT
2451 GAAACAGCCT TTGGCATGAG GGAATGGAT GTTGTGTTTT CCGTATTTCA
25 2501 AAGTTTTTTG GATGAATCAG CTTCTGGCTG GTGTTTCTTT ACTTTGGAAA
2551 ATTTAATATA TGGTATAATG TTAAGAAAAT CATCTGTTAC TGTTTACCGA
2601 TGGCAGGGGA TTTTATTCC AGTTGAGGAT TAAATATAG AAAATCCTAA
2651 AACTTGTGAG GCCTTTAATA TTGGTTTTTC TCCCTACTTT GTGATTACTC
2701 ATGAAGAAAG AAATGAAGAA AAGCCTTCTC TTAACAGTGT GTTTACATTCT
30 2751 ACATCTGGAT TTAATTTATT CCTGGTACAA ACAATCATT TCTGGAAAG
2801 TTCTCAAGTA AGATATTTTA CTTCAGACAG CCAAGATTAT TTAATCATTG
2851 CAAGTCAAAG AGATGATTCC GAATTAATC AGGTCTTCAG GTGGAATGGA
2901 GGAAGCTTCG TGTTGCATCA AAAACTCCCT GTCCGAGGTG TGCTGACCGT
2951 GGCTTGTTT AACAAGGGAG GCTCTGTGTT CTTAGCCATT TCCCAGGCTA
35 3001 ATGCCAGGCT AAACCTCCCT TTATTGAGAT GGTCTGGCAG TGGGTTTTATT
3051 AACTTTCAAG AGGTGCCTGT CAGTGGGACA ACAGAAGTTG AGGCTTTGTC
3101 TTCAGCCAAT GATATTTACC TAATATTTGC CAAAAATGTC TTTCTAGGAG
3151 ATCAGAATT CATTGATATT TTCTCTGGG AGATGGGACA GTCTTCCTTC
3201 AGGTATTTTC AGTCTGTAGA TTTTGCTGCT GTTAACAGAA TCCACTCCTT
40 3251 CACACCAGCC TCAGGAATAG CCCACATACT TCTTATTGGC CAAGATATGT
3301 CTGCTCTTTA CTGCTGGAAT TCGGAGCGTA ATCAATTCTC TTTTGTCTG
3351 GAAGTACCTT CTGCTTATGA TGTGGCTTCT GTTACAGTAA AGTCCCTTAA
3401 TTCAAGCAAG AATTTAATAG CTCTAGTGGG AGCTCATTCA CATATATATG
3451 AGCTAGCCTA CATTTCAGC CATTCTGACT TTATTCCTAG TTCAGGTGAA
45 3501 CTGATATTTG AACCTGGTGA GAGAGAAGCT ACAATAGCAG TAAATATCCT
3551 TGATGATACA GTTCCAGAAA AAGAAGAATC CTTCAAAGTT CAACCTAAAA
3601 ATCCCAAAGG AGGAGCAGAG ATTGGCATT TATGATTCTGT AACAATAACC
3651 ATTCTGTCTA ATGATGATGC CTATGGAATT GTTGCATTTG CTCAGAATTC
3701 ATTATATAAG CAAGTGGAAG AAATGGAGCA AGATAGCCTA GTAACCTTGA
50 3751 ACGTTGAACG CTTAAAAGGA ACATATGGCC GTATAACCAT AGCATGGGAA
3801 GCTGATGGAA GTATTAGTGA TATATTTCTT ACCTCAGGAG TGATTTTATT
3851 TACTGAAGGC CAGGTACTGT CAACAATCAC TCTAACTATT CTTGCTGATA
3901 ATATACCAGA GTTATCAGAG GTTGTGATTG TAACCTCAC CCGTATCACC
3951 ACAGAAGGGG TTGAGGACTC ATACAAAGGT GCTACTATTG ATCAGGACAG
55 4001 AAGCAAGTCT GTTATAACAA CTTTGCCCAA TGACTCACCT TTTGGCTTGG
4051 TGGGCTGGCG TGCTGCTCT GTCTTCATTA GAGTAGAGA GCCTAAAGAA
4101 AACACCACCA CTCTTCAGTT ACAAATAGCT CGAGATAAAG GACTACTTGG
4151 GGATATTGCC ATTCACTTGA GAGCTCAACC CAATTTCTTA CTGCATGTCG

```

4201 ATAATCAAGC TACTGAGAAT GAAGATTATG TATTGCAAGA AACATAATA
4251 ATAATGAAAG AAAACATAAA AGAAGCTCAT GCCGAAGTTT CCATTTTGCC
4301 GGATGACCTT CCTGAATTGG AGGAAGGATT TATTGTCACT ATCACTGAGG
4351 TGAACCTGGT GAACTCTGAC TTCTCTACAG GACAGCCAAG TGTGCGGAGG
5 4401 CCCGGAATGG AAATAGCTGA GATAATGATA GAAGAAAATG ACGATCCCAG
4451 AGGAATTTTT ATGTTTCATG TTACTAGAGG CGCTGGGGAA GTTATTACTG
4501 CCTATGAGGT GCCTCCACCC TTGAACGTTT TTCAAGTTCC TGTAGTCCGG
4551 CTGGCTGGAA GCTTTGGGGC AGTAAATGTT TATTGGAAAG CATCACCAGA
4601 CAGTGCTGGC CTGGAAGACT TTAACCATC TCATGGGATT CTTGAATTTG
10 4651 CAGATAAACA GGTACTGCA ATGATAGAAA TCACCATAAT TGATGATGCT
4701 GAATTTGAAT TGACAGAGAC GTTCAATATT TCCTTGATCA GTGTTGCTGG
4751 AGGTGGCAGA CTTGGTGATG ATGTTGTTGGT AACTGTTGTT ATTCCACAAA
4801 ATGATTCTCC ATTTGGAGTA TTTGGATTTG AAGAAAAGAC TGTAAGTTAA
4851 ACATATCAGG GGAAAGCCTT GTTTCAGGCT AGCGTTTCAT GTAATTTTGA
15 4901 GTAGAAAAGT TCTCACATTT TTGTTTTGGA AGTCTTGGCC AGGCATGGTG
4951 GCTCATGCCA GTAATCCCAG CACTTTGGGA GGCCGCAGCG GGCAGATCAC
5001 GAGGTCAGGA GATTGACACC ATCCTGGCCA ATATGGTTGA ATTCCCGTCT
5051 CTACTGAAAG TACAAAAATT AGCTGGGCGT GGTGGCACAT GCCTGTATTC
5101 CCAGATACTT GGGAGGCTGA GGCAGGAGAC TCGCTTGAAC CCAGGAGGCA
20 5151 GAGGTTGCAG TGAGCTGAGA TCACGCCATT GCACTCCAGC CTGGCGACAT
5201 AGAGAGACTC CATCTCAAAA AAAAAAAAAA AAAAAG

```

BLAST Results

25

No BLAST result

30

Medline entries

No Medline entry

35

Peptide information for frame 3

```

40 ORF from 0 bp to 4847 bp; peptide length: 1616
Category: putative protein
Classification: Cell signaling/communication
Prosites motifs: MULTICOPPER_OXIDASE1 (151-171)

```

45

```

1 DAWADAWALY TCATLCLKER ACSAFSFFSA SEGPQCFWMT SWISPAVNNS
51 DFWTYRKNT RVASLFSGQA VAGSDYEPVT RQWAIMQEGD EFANLTVSIL
101 PDDFPMDDES FLISLLEVHL MNISASLKNQ PTIGQPNIST VVIALNGDAF
151 GVFIYISISP NTSEDGLFVE VQEQPQTLVE LMIHRTGGSL GQVAVEWRVV
50 201 GGTATEGLDF IGAGEILTFA EGETKKTVIL TILDDSEPED DESIIVSLVY
251 TEGGSRI LPS SDTVRVNILA NDNVAGIVSF QTASRSVIGH EGEILQFHVI
301 RTFPGRGNVT VNWKIIGQNL ELNFANFSGQ LFFPEGSLNT TLFVHLLDDN
351 IPEEKEVYQV ILYDVRTQGV PPAGIAL LDA QGYAAVLTV ASDEPHGVNL
401 FALSSRFVLL QEANITIQLF INREFGSLGA INVTYTTVPG MSLKLNQTVG
55 451 NLAEPVDFV PIIGFLILEE GETAAAINIT ILEDDVPELE EYFLVNLTYV
501 GLTMAASTSF PPRLDSEGLT AQVIIDANDG ARGVIEWQQS RFEVNETHGS
551 LTLVAQRSRE PLGHVSLFVY AQNLEAQVGL DYIFTMILH FADGERYKNV
601 NIMILDDIP EGDEKFQLIL TNPSGLELG KNTIALIIVL ANDDGPVLS

```

```

5  651 FNNSEHFFLR EPTALYVQES VAVLYIVREP AQGLFGTVTV QFIVTEVNSS
    701 NESKDLTPSK GYIVLEEGVR FKALQISAIL DTEPEMDEYF VCTLFNPTGG
    751 ARLGVHVQTL ITVLQNQAPL GLFSISAVEN RATSIDIEEA NRTVYLVNVS
    801 TNGIDLAVSV QWETVSETAF GMRGMDVVFS VFQSFLESA SGWCFFTLN
10  851 LIYGIMLRKS SVTVYRWQGI FIPVEDLNIE NPKTCEAFNI GFSPYFVITH
    901 EERNEEKPSL NSVFTFTSGF KLFLVQTIIE LESSQVRYFT SDSQDYLIIA
    951 SQRDDSELTQ VFRWNGGSFV LHQKLPVRGV LTVALFNKGG SVFLAISQAN
    1001 ARLNSLLFRW SGSGFINFQE VPVSGTTEVE ALSSANDIYL IFAKNVFLGD
    1051 QNSIDIFIWE MGQSSFYRFQ SVDFAAVNRI HSFTPASGIA HILLIGQDMS
15  1101 ALYCWNSESN QFSFVLEVPS AYDVASVTVK SLNSSKNLIA LVGAHSHIYE
    1151 LAYISSHSDF IPSSGELIFE PGEREATIAV NILDDTVPEK EESFKVQLKN
    1201 PKGGAEIGIN DSVTITILSN DDAYGIVAFQ QNSLYKQVEE MEQDSLVTLN
    1251 VERLKGTYGR ITIAWEADGS ISDIFPTSGV ILFTEGQVLS TITLTILADN
    1301 IPELSEVVIV TLTRITTEGV EDSYKGATID QDRSKSVITT LPNDSPFGLV
15  1351 GWRAASVFIR VAEPKENTTT LQLQIARDKG LLGDIAIHLR AQPNFLLHVD
    1401 NQATENEDYV LQETIIMKE NIKEAHAEVS ILPDDLPELE EGFIVTITEV
    1451 NLVNSDFSTG QPSVRRPGME IAEIMIEEND DPRGIFMFHV TRGAGEVITA
    1501 YEVPPPLNVL QVPVVRLAGS FGAVNVYWKA SPDSAGLEDF KPSHGILEFA
    1551 DKQVTAMIEI TIIDDAEFEL TETFNISLIS VAGGGRLGDD VVVTVVIPQN
20  1601 DSPFGVFGFE EKTVS

```

BLASTP hits

25

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_10p7, frame 3

```

30  TREMBL:AF055084_1 gene: "VLGR1"; product: "very large G-protein
    coupled
    receptor-1"; Homo sapiens very large G-protein coupled receptor-
    1
    (VLGR1) mRNA, complete cds., N = 3, Score = 284, P = 1.2e-33
35

```

```

    TREMBL:DMAF9897_1 gene: "Calx"; product: "CALX"; Drosophila
    melanogaster 3Na(+)-1Ca(2+) exchanger (Calx) mRNA, complete cds.,
    N =
    1, Score = 178, P = 3.3e-09
40

```

```

    >TREMBL:AF055084_1 gene: "VLGR1"; product: "very large G-protein
    coupled
    receptor-1"; Homo sapiens very large G-protein coupled
45  receptor-1 (VLGR1)
    mRNA, complete cds.
    Length = 1,967

```

HSPs:

50

```

    Score = 284 (42.6 bits), Expect = 1.2e-33, Sum P(3) = 1.2e-33
    Identities = 192/738 (26%), Positives = 314/738 (42%)

```

Query: b7

```

55  SGQAVAGSDYEPVTRQWAIMQEGDEFANLTVSILPDDFPemdeSFLISLLEVHLMNISAS 126
      S + G DY + Q G + + +SI+ D+ E +E +E+
L +

```

Sbjct: 102 SSASPGGVYI-LHGSTVTFQHGQNLSEFINISIIDDNESEFEFEP-----
IEILLTGATGG 155

Query: 127

5 LKNQPTIGQPNISTVVIALNGDAFGVFVIYSISPNTSEGLFVEVQEQPQTLV-ELMIHR 185
+G+ +S ++IA + FGV N S+ + +

T++ L++ R

Sbjct: 156 A----VLGRHLVSRIIIAKSDSPFGVIRFL----NQSK----
ISIANPNSTMILSLVLER 203

10

Query: 186 TGGSLGQVAVEWRVVGGTATEGL-----DFIG-AGEILTFAEGETK-
KTVILTXXXXXXX 238

TGG LG++ V W VG + E L D + F EGE

+T+ILT

15

Sbjct: 204
TGGLLGEIQVNWETVGPNSQEALLPQNRDIADPVSGLFYFGE GEGGVRTIILTIYPHEEI 263

Query: 239 XXXXXXXXXXXLVYTEGGSRIPLSSDTVRVNILANDNVAGIVSF--
QTASRSVIGH-----EG 292

20

L +G +++ + V + I + G+V F +T S+

EG

Sbjct: 264

EVEETFIIKLHLVKGEAKLDSRAKDVTLTIQEFDPNGVVQFAPETLSKKTYSEPLALEG 323

25

Query: 293 EILQFHVIRTFPGR-GNVTVNWKIIGQ-
NLELNANFANFSGQLFFPEGSLNTTLFVHLLDDN 350

+L +R G G + V W++ + ++ +F + SG +G +

VHLL D

Sbjct: 324

30

PLLITFFVRRVKGTGFEIMVYWELSSEFDITEDFLSTSGFFTADGESEASFVHLLPDE 383

Query: 351

IPEEKEVYQVILYDVRTQGVPPAGIALLDAGGYAAVLTVEASDEPHGVNLNLFAL-SSRFVL 409
+PE +E Y + L V G A LD + +V A+D+PHGV

35

FAL S R +

Sbjct: 384 VPEIEEDYVIQLVSVSE-----GGAELDLEKSITWFSVYANDDPHGV--
FALYSRQSI 434

Query: 410 LQEANI--TIQLFINREFGSLGAINVTYTTVPGMLSLKNQT-
VGNLAEPEVDVPIIGFL 466

40

L N+ +IQ+ I R G+ G + V K Q V AE +

L

Sbjct: 435 LIGQNLIRSIQINITRLAGTFGDVAVGLRISSDH---KEQPIVTENAERQ--
-----L 482

45

Query: 467

ILEEGETAAAINITILEDDVPELEEYFLVNLTYYVGLTMAASTSFPPRLDSEGLTAQVIID 526
++++G T + I L F + L V L P L E

+A V+

50

Sbjct: 483 VVKDGATYKVDVPIKNQVFLSLGSNFTLQLVTVMVLVGGRFYGMPTILQ-
EAKSA-VLPV 540

Query: 527 ANDGARGVIEWQSRFEV-NETHGSLTLVAQRSREPLGHVSLFV---
YAQNLEAQVGLDY 582

55

+ A + ++ + F++ N T G+ ++ R R G +S+ YA

LE +

Sbjct: 541 SEKAANSQVGFESTAFQLMNITAGTSHVMISR-
RGTYGALSVAWTTGYAPGLEIPEFIVV 599

Query: 583 -IFTPMI--
 LHFADGERYKNVNIMILDDDIPEGDEKFQILITNPSPGLELGKNTIALIIV 639
 TP + L F+ GE+ K V + P E F L L+ G

5 + IV
 Sbjct: 600 GNMTPTLGSLSFSGEQRKGVFLWTFPS--
 PGWPEAFVLHLSGVQSSAPGGAQLRSGFIV 657

Query: 640 LANDDGPVLSFN-
 10 NSEHFFLREPTALYVQESVAVLYIVREPAQGLFGTVTVQFIVTEVN 698
 A + GV F+ +S + + E T + ++ V L+ G +
 T
 Sbjct: 658 -AEIPEMGVFQFSTSSRNIIIVSEDTQM-IRLHVQRLF-----
 GFHSDLIKVSQTTAG 708

15 Query: 699 SSNESKDLTP-SKGYIVLEEGVRFKALQISAILDTEPEMDEYFVCTL-----
 -----FNP 747
 S+ +D P G + ++ +I+ I D E++E+F L
 F+

20 Sbjct: 709
 SAKPLEDFEPVQNGELFFQKFQTEVD FEITIINDQLSEIEEFFYINLTSVEIRGLQKFDV 768

Query: 748 TGGARLGVHVQT-LITVLQNQAPLGLFSISAVENR-ATSIDIE----
 EANRTVYLNVSRT 801

25 RL + +IT+L N G+ IS E A ++D E T
 YL+ S+T
 Sbjct: 769 NWSPRNLDFSVAVITILDNDLAGM-
 DISFPETTAVAVDITLIPVETESTTYLSTSKT 827

30 Query: 802 NGI 804
 I
 Sbjct: 828 TTI 830

Score = 266 (39.9 bits), Expect = 4.0e-25, Sum P(3) = 4.0e-25
 35 Identities = 175/708 (24%), Positives = 306/708 (43%)

Query: 131
 PTIGQPNISTVVIALNGDAFGVFIYSSISPNTSEDGLFVEVQEQPQTLVELMIHRTGGSL 190
 P IG +I ++I N +A G+ P + EV+E L+ +

40 + R G+
 Sbjct: 39 PEIGNISIVRIIIMKNDNAEGII---EFDPKYTA----FEVEEDVG-
 LIMIPVVRHLGTY 90

Query: 191 GQVAVEWRVVGGTATEG-
 45 LDFIGAGEILTFAEGETKKTIVILTXXXXXXXXXXXXXXXXXXLV 249
 G V ++ +A+ G +D+I G +TF G+ + ++
 L
 Sbjct: 91
 GYVTADFISQSSASPGGVYILHGSTVTFQHGQNLFINISIIDNESEFEEPIEILLT 150

50 Query: 250 YTEGGSRIPLSSDTPVRVNILANDNVAGIVSFQATASRSVIGHEGE--
 ILQFHVIRTFPGRG 307
 GG+ +L R+ I +D+ G++ F S+ I + IL +

RT G
 55 Sbjct: 151 GATGGA-
 VLGRHLVSRIIIAKSDSPFGVIRFLNQSKISIANPNSTMILSLVLERTGGLLG 209

Query: 308 NVTVNWKIIGQN-----LELN--FAN-FSGQLFFPEGSLNT-
 TLFVHLLDDNIPEEKEY 358
 + VNW+ +G N L N A+ SG +F EG T+ + +
 E +E +
 5 Sbjct: 210
 EIQVNWETVGPNSQEALLPQNRDIADPVSGLFYFGE GEGGVRTIILTIYPHEEIEVEETF 269

Query: 359 QVILYDVRTQGVPPAGIALLDARQYAAVLTVEASDEPHGVLNFA---
 LSSRFV---LLQE 412
 10 + L+ V+ G A LD++ LT++ +P+GV+ FA LS +
 L E
 Sbjct: 270 IIKLHLVK-----
 GEAKLDSRAKDVTLTIQEFQDPNGVVQFAPETLSKKTYS EPLALE 322

Query: 413
 ANITIQLFINREFGSLGAINVTYTTVPGLSLKNQTVGNLAEPEVDFVPIIGFLILEEGE 472
 + I F+ R G+ G I V + L ++ ++ E DF+
 GF + +GE
 Sbjct: 323 GPLLITFFVRRVKGTGFEIMVYW-----ELSEF--DITE---
 20 DFLSTSGFFTADGE 370

Query: 473
 TAAAINITILEDDVPELEEYFLVNLTYVGLTMAASTSFPPRLDSEGLTAQVIIDANDGAR 532
 + A+ ++ +L D+VPE+EE +++ L S LD E
 25 + AND
 Sbjct: 371 SEASFDVHLLPDEVPEIEEDYVIQLV-----
 SVEGGAELDLEKSITWFSVYANDDPH 422

Query: 533 GVIEWQQRFEV---NETHGSLTLVAQRSREPLGHVS--
 30 LFFVAQNLEAQVGLDYIFTPM 587
 GV R + S+ + R G V+ L + + + E +
 + +
 Sbjct: 423
 35 GVFALYSDRQSIILIGQNLIRSIQINITRLAGTFQDVAVGLRISSDHKEQPIVTENAERQL 482

Query: 588 ILHFADGERYKNVNIMILDDDI--PEGDE-KFQLILTNPSPGLELGKNTI--
 -ALIIVLA 641
 ++ DG YK V+++ + + + G QL+ G G TI
 A VL
 40 Sbjct: 483 VVK--DGATYK-
 VDVVPIKNQVFLSLGSNFTLQLVTVMLVGGRFYGMPTILQEA KSAVLP 539

Query: 642
 NDDGPGVLSFNNSEHFFLREPTALYVQESVAVLYIVREPAQGLFGTVTVQFIV-----TE 696
 45 + NS+ F E TA + A V +G +G ++V +
 E
 Sbjct: 540 VSEKAA-----NSQVGF--
 ESTAFQLMNITAGTSHVMISRRGTYGALSVAWTTGYAPGLE 592

Query: 697
 VNSSNESKDLTPSKGYIVLEEGVRFKALQISAILDTEPEMDEYFVCTLFNPTGGARLG VH 756
 + ++TP+ G + G + K + + P E FV L
 A G
 Sbjct: 593 IPEFIVVGNMTPTLGSLSFSHGQARKGVFLWTF--
 55 PSPGWPEAFVLHLSGVQSSAPGGAQ 650

Query: 757 VQTLITVLQNAQPLGLFSISAVENRATSIDIEEANRTVYLVNVSRTNGI--
 DLAVSVQWET 814

+++ V + + P+G+F S +R +I + E + + L+V R G DL
+ V ++T
Sbjct: 651 LRSGFIVAEIE-PMGVFQFST-SSR--
NIIVSEDTQMIRLHVQRLFGFHSDL-IKVSQT 705

5 Query: 815 VSETAFGMRGMDVVFS---VFQSFLE 838
+ +A + + V + FQ F E
Sbjct: 706 TAGSAKPLEDFEPVQNGELFFQKFQTE 732

10 Score = 246 (36.9 bits), Expect = 4.1e-32, Sum P(3) = 4.1e-32
Identities = 92/338 (27%), Positives = 157/338 (46%)

Query: 511 PPRLDSEGLTAQVIIDANDGARGVIEW--
QQSRFEVNETHGSLTLVAQRSREPLGHVSLF 568

15 PP + + + ++II ND A G+IE+ + + FEV E G + + R
G+V+
Sbjct: 38 PPEIGNISIV-
RIIMKNDNAEGIIIEFDPKYTAFEVEEDVGLIMIPVVRHLHGTYGYVTAD 96

20 Query: 569 VYAQNLEAQVG-
LDYIFTPMILHFADGERYKNVNIMILDDDIPEGDEKFLILTNPSPGL 627
+Q+ A G +DYI + F G+ +NI I+DD+ E +E
+++LT + G
Sbjct: 97
FISQSSASPGGVYILHGSTVTFQHGQNLFINISIIDNESEFEEPIEILLTGATGGA 156

25 Query: 628
ELGKNTIALIIVLANDDGPVLSFNNSEHFFLREPTALYVQESVAVLYIVREPAQGLFGT 687
LG++ ++ II+ +D GV+ F N + P S +L +V E
GL G
Sbjct: 157 VLGRHLVSRIIIAKSDSPFGVIRFLNQSKISIANPN-----
STMILSLVLERTGGLLGE 210

30 Query: 688 VTVQFIVTEVNSSN----ESKDLT-PSKGYIVLEEGVR-
FKALQISAILDTEPEMDEYFV 741
+ V + NS +++D+ P G EG + + ++ E
E++E F+
Sbjct: 211
IQVNWETVGPNSQEAALLPQNRDIADPVSGLFYFGE GEGGVRTIILTIYPHEEIEVEETFI 270

40 Query: 742 CTLFNPTGGARLGHVHQT-ITVLQNAAPLGL--FSISAVENRATSIDIE-
EANRTVYLN 797
L G A+L + + +T+ + P G+ F+ + + S + E
+
Sbjct: 271
IKLHLVKGEAKLDSRAKDVTLTIQEFGDPNGVVQFAPETLSKKTYSEPLALEGPLLITFF 330

45 Query: 798 VSRTNGIDLAVSVQWETVSETAFGMRGMDVVFSVFQSFLEDESASGWCFRTL
848
V R G + V WE SE F + + FL S SG FFT+
Sbjct: 331 VRRVKGTFGIMVYWELSSE-----FDITEDFL--STSG--FFT
366

50 Score = 246 (36.9 bits), Expect = 1.9e-19, Sum P(3) = 1.9e-19
Identities = 87/303 (28%), Positives = 138/303 (45%)

55 Query: 1162 PSSGELIFEPGEREA-TIAVNILDDTVPEKEESFKVQLKNPKGGAEIGIN-
DSVTITILS 1219

P SG F GE TI + I E EE+F ++L KG A++

VT+TI

Sbjct: 236

PVSGLFYFGE GEGGVRTIILTIYPHEEIEVEETFIKHLVKGEAKLDSRAKDVTLTIQE 295

5

Query: 1220 NDDAYGIVAFANSL----

YKQVEEMEQDSLVTNLVERLKGTYGRITIAWEADGSIS--- 1272

D G+V FA +L Y + +E L+T V R+KGT+G I + WE

Sbjct: 296

10 FGD PNGVVQFAPETLSKKTYS EPLALEGPLLITFFVRRVKGTFG EIMVYWELSS EFDITE 355

Query: 1273

DIFPTSGVILFTEGQVLSTITLTILADNIPELSEVVIVTLTRITTEGVEDSYKGATIDQD 1332

D TSG +G+ ++ + +L D +PE+ E ++ L ++ EG

15 GA +D +

Sbjct: 356 DFLSTSGFFTIADGESEASF DVHLLPDEVPEIEEDYVIQL--VSVEG-----

-GAELDLE 407

Query: 1333

20 RSKSVITTLPNDS PFGLVGWRAASVFIRVAEPKENTTTTLQLQIARDKGLLGDI AHLRAQ 1392

+S + + ND P G+ + I + + ++Q+ I R G

GD+A+ LR

Sbjct: 408 KSITWFSVYANDDPHGVFALYSDRQ SILIGQ--

NLIRSIQINITRLAGTFGDVAVGLRIS 465

25

Query: 1393 PNFL LHVDNQ-

ATENEDYVLQETIIMKENIKEAHA EVSILPDDLPELEEGFIVTITEVN 1451

+ H + TEN E +++K+ V I L F

+ + V

30

Sbjct: 466 SD---HKEQPIVTENA-----

ERQLVVKDGATYKVDVPIKNQVFLSLGSNFTLQLVTVM 517

Query: 1452 LVNSDFSTGQPSV 1464

LV F G P++

35

Sbjct: 518 LVGGRFY-GMPTI 529

Score = 246 (36.9 bits), Expect = 1.9e-19, Sum P(3) = 1.9e-19

Identities = 89/334 (26%), Positives = 150/334 (44%)

40

Query: 1159

DFIPSSGELIFEPGEREATIAVNILDDTVPEKEESFKVQLKNPKGGAEIGINDSVTITIL 1218

D+I + F+ G+ + I ++I+DD E EE ++ L GGA +G +

I I

Sbjct: 110

45

DYILHGSTVTFQHGQNL SFINISIIDNESEFEEPIELLTGATGGAVLGRHLVSRIIIA 169

Query: 1219

SNDDAYGIVAFANSLYKQVEEMEQDSLVTNLVERLKGTYGRITIAWEADGSIS----- 1272

+D +G++ F S + +++L +ER G G I + WE G

50

S

Sbjct: 170 KSDSPFGVIRFLNQSKIS-

IANPNSTMILSLVLERTGGLLGEIQVNWETVGPNSQEALLP 228

Query: 1273 ---DIF-PTSGVILFTEGQV-

55

LSTITLTILADNIPELSEVVIVTLTRITTEGVEDSYKGA 1327

DI P SG+ F EG+ + TI LTI E+ E I+ L + E

DS

Sbjct: 229

QNRDIADPVSGLFYFGE GEGGVRTIILTIYPHEEIEVEETFIKHLVKGEAKLDS---- 284

5 Query: 1328 TIDQDRSKSVITTLPN-DSPFGLVGWRAASVFIRV-AEPK--
ENTTTLQLQIARDKGLLG 1383

R+K V T+ P G+V + ++ + +EP E + +

R KG G

Sbjct: 285 -----

10 RAKDVTLTIQEF G DPN GVVQFAPETLSKKTYSEPLALEGPLLITFFVRRVKGTFG 339

Query: 1384

DIAIHLRAQPNFLLHVDNQATENEDYVLQETIIMKENIKEAHAEVSIPLDDLPELEEGF 1443

+I ++ F + ED++ + + EA +V

+LPD++PE+EE +

15 Sbjct: 340 EIMVYWELSSSEFDI-----

TEDFLSTSGFFTIADGESEASFVHLLPDEVPEIEEDY 391

Query: 1444 IVTITEVNLVNSDFSTGQPSVRRPGMEIAEIMIEENDDPRGIFMFHVTR
1492

++ + V + + + I + NDDP G+F + R

20 Sbjct: 392 VIQLVSVE-----GGAELDLEK---SITWFSVYANDDPHG VFALYSDR
431

25 Score = 237 (35.6 bits), Expect = 9.4e-34, Sum P(3) = 9.4e-34
Identities = 101/367 (27%), Positives = 165/367 (44%)

Query: 67

SGQAVAGSDYEPVTRQWAIMQEGDEFANLTVSILPDDFPEDDESFLISLLEVHLMNISAS 126

S + G DY + Q G + + +SI+ D+ E +E +E+

30 L +

Sbjct: 102 SSASPGGVYDI-LHGSTVTFQHGQNL SFINISIIDNESEFEPEP-----

IEILLTGATGG 155

Query: 127

35 LKNQPTIGQPNISTVVIALNGDAFGVFVIYSISPNTSEDGLFVEVQEQPQTLVELMIHRT 186

+G+ +S ++IA + FGV N S+ +

++ L++ RT

Sbjct: 156 A----VLGRHLVSRIII AKSDSPFGVIRFL----NQSKISI---

ANPNSTMILSLVLERT 204

40

Query: 187 GGS LGQVAVEWRVVGGTATEGL-----DFIG-AGEILTFAEGETK-
KTVILTXXXXXXXXX 239

GG LG++ V W VG + E L D + F EGE +T+ILT

Sbjct: 205

45 GGLLGEIQVNWETVGPNSQEAALLPQNRDIADPVSGLFYFGE GEGGVRTIILTIYPHEEIE 264

Query: 240 XXXXXXXXLVYTEGGSRILPSSD TVRVNLANDNVAGIVSF--

QTASRSVIGH----EGE 293

L +G +++ + V + I + G+V F +T S+

50 EG

Sbjct: 265

VEETFIKHLVKGEAKLDSRAKDVTLTIQEF G DPN GVVQFAPETLSKKTYSEPLALEGP 324

Query: 294 ILQFHVIRTFPGR-GNVTVNWKIIGQ-

55 NLELN FANFSGQLFFPEGSLNTTLFVHLLDDNI 351

+L +R G G + V W++ + ++ +F + SG +G. +

VHLL D +

Sbjct: 325

LLITFFVRRVKGTGTFGEIMVYWELSSEFDITEDFLSTSGFFFTIADGESEASFDVHLLPDEV 384

Query: 352

5 PEEKEVYQVILYDVRTQGVPPAGIALLDAGGYAAVLTVEASDEPHGVNLFAL-SSRFVLL 410
 PE +E Y + L V G A LD + +V A+D+PHGV FAL

S R +L

Sbjct: 385 PEIEEDYVIQLVSV-----GGAELDLEKSITWFSVYANDDPHGV--
 FALYSDRQISIL 435

10

Query: 411 QEANI--TIQLFINREFGSLGAINV 433

N+ +IQ+ I R G+ G + V

Sbjct: 436 IGQNLIRSIQINITRLAGTFGDVAV 460

15

Score = 230 (34.5 bits), Expect = 2.3e-14, Sum P(3) = 2.3e-14
 Identities = 98/368 (26%), Positives = 164/368 (44%)

Query: 1240 EMEQD-

20 SLVTLNVERLKGTYGRITIAWEADGSISDIFPTSGVILFTEGQVLSTITLILA 1298
 E+E+D L+ + V RL GTYG +T + + S + P GV G

ST+T

Sbjct: 71 EVEEDVGLIMIPVVRHLHGTYGYVTADFISQSSSAS--P-GGVYILHG---
 STVTFQH-G 123

25

Query: 1299 DNIPELSEVVIVTLTRITTEGVEDSYKGATIDQDRSKSVITTL---
 PNDSPFGLVGWRAA 1355

N+ ++ +I E +E GAT + +++ +

+DSPFG++ +

Sbjct: 124

30 QNLSFINISIIDDNESEFEPIEILLTGATGGAVLGRHLVSRIIIKSDSPFGVIRFLNQ 183

Query: 1356 SVFIRVAEPKENTTTLLQIARDKGLLGDIHLRAQ-
 PNFLHVDNQATENEDYVLQET 1414

S I +A P +T L L + R GLLG+I ++ PN + Q +

35

D V

Sbjct: 184 SK-ISIANPN-

STMILSLVLERTGGLLGEIQVNWETVGPNSQEALLPQNRDIADPV--SG 239

40

Query: 1415 IIIMKENIKEAHAEV-

SILPDDLPELEEGFIVTITEVNLVNSDFSTGQPSVRRPGMEIAE 1473

+ E + +I P + E+EE FI+ +++LV G+ +

++

Sbjct: 240 LFYFGEGEVVRTIILTIYPHEEIEVEETFII---KLHLVK-----
 GEAKLDSRAKDV- 290

45

Query: 1474

IMIEENDDPRGIFMFHVTRGAGEVITAYEXXXXXXXXXXXXXXXXXXAGSFGAVNVYWKASPD 1533

+ I+E DP G+ F + + + G+FG +

VYW+ S +

50

Sbjct: 291

LTIQEFGDPNGVVFAPETLSKKTYSPLALEGPLLITFFVRRVKGTGTFGEIMVYWELSSE 350

Query: 1534

55 SAGLEDFKPSHGILEFADKQVTAMIEITIIDDAEFELTETFNISLISVAGGGRLGDDVVV 1593

EDF + G AD + A ++ ++ D E+ E + I L+SV GG

L + +

Sbjct: 351

FDITEDFLSTSGFFFTIADGESEASFDVHLLPDEVPEIEEDYVIQLVSVEGGAELDLEKSI 410

Query: 1594 T-VVIPQNDSPFGVF 1607

T + ND P GVF

Sbjct: 411 TWFSVYANDDPHGVF 425

5

Score = 190 (28.5 bits), Expect = 7.5e-11, Sum P(3) = 7.5e-11
Identities = 136/591 (23%), Positives = 247/591 (41%)

Query: 67

10 SGQAVAGSDYEPVTRQWAIMQEGDEFANLTVSILPDDFPEMDESFLISLLEVHLMNISAS 126
+G A D+EPV Q+ + ++I+ D E++E F I+L V

+ +

Sbjct: 707

AGSAKPLEDFEPVQNGELFFQKFQTEVD FEITIINDQLSEIEEFFYINLTSVEIRGLQKF 766

15

Query: 127 LKN-QPTIGQP-NISTVVIALNGDAFGVFVIY-
SISPNTSEDGLFVEVQEQPQTLVELMI 183

N P + +++ + I N D G+ + + + + D + V+ +

T L

20

Sbjct: 767

DVNWSPRLNLD FSVAVITILDND DLAGMDISFPETTAVAVD TTLPVETESTTY--LST 824

Query: 184 HRTGGS LGQVAVEWRVVG GTATEGLDFIGAGEILTF--
AEGETKKT VILTXXXXXXXXXXXX 241

25

+T L V +V T G+ I +++T ++K + T

Sbjct: 825 SKTTTILQPTNVV-AIV--TEATGVSAIPE-
KLVTLHGTPAVSEKPDVATVTANVSIHGT 880

30

Query: 242 XXXXX XLVYTEGGSRLPSSD TVRVN ILANDNVAGIVSF--
QTASRSVIGHEGEILQFHV 299

+VY E + + +T V I G VS +T E

L F

Sbjct: 881 FSLGPSIVYIEEEMKN-

GTFN TAEVLIRRTGGFTGNVSITVKT FGERCAQMEPNALPF-- 937

35

Query: 300

IRTFPGRGNVT VNWKIIGQNL ELNFANFSGQLFFPEGSLNTTLFVHLLDDNIPEEKEVYQ 359
R G N+T W + E +F + L F +G + V +LDD+

PE +E +

40

Sbjct: 938 -RGIYGISNLT--WAVE----

EEDFEEQTLTLIFLDGERERKVS VQILDDDEPEGQEFFY 990

Query: 360 VILYDVRTQGVPPAGIAL LDAQ---GYAA--
VLTVEASDEPHGVLN FALSSRFVL-LQEA 413

45

V L + P G +++ + G+AA ++ + SD +G++ F+ S+

L L+E

Sbjct: 991 VFLTN-----

PQGGQIIVEGKDDTGFAAFAMVIITGSDLHNGIIGFSEESQSGLELREG 1044

50

Query: 414 NITIQ LFI-----NREFGSLGAI-
NVTYTTVP GMLSLKNQTVGNLAEPEVDFVPIIGFL 466

+ +L + NR F + VT ++ L+ V NL E E+

V G

55

Sbjct: 1045 AVMRRLHLIVTRQPNRAFEDVKVFWRVTLNKT--VVVLQKDG V-NLME-
ELQSVS--GTT 1098

Query: 467 ILEEGETAAAINITILEDDVPELEEFVNL--
TYVGLTMAASTSFPPRLDSEGLTAQVI 524

G+T I+I + + VP++E YF V L

G + S F

E +Q +

Sbjct: 1099

TCTMGQTKCFISIELKPEKVPQVEVYFFVELYEATAGAAINNSARFAQIKILESDESQSL 1158

5

Query: 525 IDANDGARGVIEWQQRSF---EVNETHGS-

LTLVAQRSREPLGHVSLFVYAQNLEAQVGL 580

+ + G+R + +++ +V G+ L + S + L

A G

10 Sbjct: 1159

VYFSVGSRLAVAHKKATLISLQVARDSGTGLMMSVNFSTQELRSAETIGRTIISPAISGK 1218

Query: 581 DYIFTPMILHFADGERYKNVNIMILDD--

DIPEGDEKFQILITNPSPGLELGKNTIALII 638

15

D++ T L F G+R +++++ + + ++FQ++L +P G +

K I

Sbjct: 1219

DFVITEGTLVFEPGQRSTVLDVILTPETGSLNSFPKRFQIVLFDPKGGARIDKVYGTANI 1278

20 Query: 639 VLAND-DGPGVLSFNSEH 656

L +D D + + H

Sbjct: 1279 TLVSDADSQAIWGLADQLH 1297

Score = 188 (28.2 bits), Expect = 1.2e-33, Sum P(3) = 1.2e-33
 25 Identities = 84/329 (25%), Positives = 146/329 (44%)

Query: 1126 SVTVKSLNS-----

SKNLIALVGAHSHIYELAYISSHSDFIPSSGELIFEPGEREATIAV 1180

S+TVK+ N + G + I L + DF + LIF

30 GERE ++V

Sbjct: 917 SITVKTGGERCAQMEPNALPFRGIYG-

ISNLTWAVEEEDFEEQTLTLIFLDGERERKVS 975

Query: 1181 NILDDTVPEKEESFKVQLKNPKGGAEI--GINDS----VTITILSNDDAY-
 35 GIVAFQNS 1233

ILDD PE +E F V L NP+GGA+I G +D+ + I++ D +

GI+ F++ S

Sbjct: 976

QILDDDEPEGQEFFYVFLTNPQGGAQIVEGKDDTGFAAFAMVIITGSDLHNGIIGFSEES 1035

40

Query: 1234 LYKQVEEMEQLSLVT---LNVERLKG-TYGRITIAWEAD-

GSISDIFPTSGVILFTEGQV 1288

+ E+ + +++ L V R + + + W + GV

L E Q

45 Sbjct: 1036 --

QSGLELREGAVMRRHLIVTRQPNRAFEDVKVFWRVTLNKTVVVLQKDGVLNMEELQS 1093

Query: 1289 LSTITLTILADNIPELS-EVVIVTLTRITTEGVEDSYK---

GATIDQDRSKSVITTL PND 1344

50 +S T + +S E+ + ++ + Y+ GA I+ +

I L +D

Sbjct: 1094

VSGTTTCTMGQTKCFISIELKPEKVPQVEVYFFVELYEATAGAAINNSARFAQIKILES 1153

55 Query: 1345 SPFGLVGWRAASVFIRVAEPKENTTTTLQLQIARDKG--LLGDIAI---
 HLRAQPNFLLHV 1399

LV + S R+A + T + LQ+ARD G L+ + LR+

+

Sbjct: 1154 ESQSLVYFSVGS---

RLAVAHKKATLISLQVARDSGTGLMMSVNFSTQELRSAETIGRTI 1210

Query: 1400 DNQATENEDYVLQETIIIMKENIKEAHAEVSIPLD 1434

5 + A +D+V+ E ++ + + +V + P+

Sbjct: 1211 ISPAISGKDFVITEGTLVFEPGQRSTVLDVILTPE 1245

Score = 186 (27.9 bits), Expect = 2.5e-13, Sum P(3) = 2.5e-13

Identities = 75/242 (30%), Positives = 113/242 (46%)

10

Query: 1206

EIGINDSVTITILSNDDAYGIVAFQAQNSLYKQVEEMEQDSLVTLNVERLKGTYGRITIAW 1265

EIG V I I+ ND+A GI+ F + Y E E L+ + V RL

GTYG +T +

15

Sbjct: 40 EIGNISIVRIIIMKNDNAEGIIIEF--

DPKYTAFEVEEDVGLIMIPVVRHLGTYGYVTADF 97

Query: 1266 EADGSIS-----

DIFPTSGVILFTEGQVLSTITLILADNIPELSEVVIVTLTRITTEGV 1320

20

+ S + D + F GQ LS I ++I+ DN E E + +

LT T G

Sbjct: 98

ISQSSSASPGGVVDYILHGSTVTFQHGQNLSEFINISIIDDNESEFEEPIEILLTGAT--G- 154

25

Query: 1321

EDSYKGATIDQDRSKSVITTLPNDSFGLVGWRAASVFIRVAEPKENTTTTLQQLIARDKG 1380

GA + + +I +DSPFG++ + S I +A P +T L

L + R G

30

Sbjct: 155 -----GAVLGRHLVSRIIIA-KSDSPFGVIRFLNQSK-ISIANPN-

STMILSLVLERTGG 206

Query: 1381 LLGDIAIHLRAQ-PNFLHVDNQATENEDYVLQETIIIMKENIKEAHAEV-SILPDDLPE 1438

LLG+I ++ PN + Q + D V + E +

35

+I P + E

Sbjct: 207 LLGEIQVNWETVGPNSQEAALLPQNRDIADPV--

SGLFYFGE GEGGVRTIILTIYPHEEIE 264

Query: 1439 LEEGFIVTI 1447

40

+EE FI+ +

Sbjct: 265 VEETFIKL 273

Score = 179 (26.9 bits), Expect = 9.4e-34, Sum P(3) = 9.4e-34

Identities = 65/244 (26%), Positives = 114/244 (46%)

45

Query: 581 DYIFTMILHFADGERYKNVNIMILDDDIPEGDEKFQILITNPSPGLEL--GKN-----T 633

D+ + L F DGER + V++ ILDDD PEG E F + LTNP G ++

GK+

50

Sbjct: 954

DFEEQTLTLIFLDGERERKVSQILDDDEPEGQEFFYVFLTNPQGGQIVEGKDDTGFAA 1013

Query: 634 IALIIVLANDDGPGVLSFNNSEHFFLREPTALYVQESVAVLYIVREPAQG--

---LFGTV 688

55

A++I+ +D G++ F+ L ++ L + R+P +

+F V

Sbjct: 1014 FAMVIITGSDLHNGIIGFSEESQSGLELREGAVMRR--

LHLIVTRQPNRAFEDVKVFWRV 1071

Query: 689 TVQ--
 FIVTEVNSSNESKDLTPSKGYIVLEEGVRFKALQISAILDTEPEMDEYFVCTLFN 746
 T+ +V + + N ++L G G + I + P+++

5 YF L+
 Sbjct: 1072
 TLNKTVVVLQKDGVNLMEELOSVSGTTTCTMGQTKCFISIELKPEKVPQVEVYFFVELYE 1131

Query: 747 PTGGARLGVHVQ-
 10 TLITVLQNAAPLGLFSISAVENRATSIDIEEANRTVYLVNVSRTNGID 805
 T GA + + I +L++ L S V +R ++ ++A + L
 V+R +G
 Sbjct: 1132 ATAGAAINNSARFAQIKILESDESQSLVYFS-VGSRL-AVAHKKAT-
 LISLQVARDSGTG 1188

15 Query: 806 LAVSVQWET 814
 L +SV + T
 Sbjct: 1189 LMMSVNFST 1197

20 Score = 174 (26.1 bits), Expect = 4.1e-32, Sum P(3) = 4.1e-32
 Identities = 58/200 (29%), Positives = 102/200 (51%)

Query: 1159
 25 DFIPSSGELIFEPGEREATIAVNILDDTVPEKEESFKVQLKNPKGGAEIGINDSVT-ITI 1217
 DF+ +SG GE EA+ V++L D VPE EE + +QL + +GGAE+ +
 S+T ++
 Sbjct: 356
 DFLSTSGFFTIADGESEASFDVHLLPDEVPEIEEDYVIQLVSVEGGAELDLEKSITWFSV 415

30 Query: 1218 LSNDDAYGIVAFQAQNSLYKQVEEMEQDSL--
 VTLNVERLKGTYGRITIAWEADGSISDIF 1275
 +NDD +G+ A + +Q + Q+ + + +N+ RL GT+G + +
 SD
 Sbjct: 416 YANDDPHGVFALYSD---
 35 RQILIGQNLIRSIQINITRLAGTFGDVAVGLRIS---SDHK 469

Query: 1276 PTSGVILFTEGQVLSTITLTILADNIPELSEVVI-----
 VTLTRITTEGVEDSYKGA-TI 1329
 V E Q++ T D +P ++V + TL +T V
 40 + G TI
 Sbjct: 470
 EQPIVTENAERQLVVKDGATYKVDVVPIKNQVFLSLGSNFTLQLVTVMLVGGRFYGMPTI 529

Query: 1330 DQDRSKSVITTLPNDSPPGLVGWRAAS 1356
 45 Q+ +KS + + + VG+ + +
 Sbjct: 530 LQE-AKSAVLPVSEKAANSQVGFESTA 555

Score = 145 (21.8 bits), Expect = 4.3e-24, Sum P(3) = 4.3e-24
 Identities = 104/396 (26%), Positives = 170/396 (42%)

50 Query: 88
 EGDEFANLTVSILPDDFPEMDESFLISLLEVHLMNISASLKNQPTIGQPNISTVVIALNG 147
 +G+ A+ V +LPD+ PE++E ++I L+ V A L + +I +
 + N
 55 Sbjct: 368 DGESEASFDVHLLPDEVPEIEEDYVIQLVSVEG---GAELDLEKSI-----
 TWFSVYAND 419

Query: 148

DAFGVFVIYSISPNTSEDGLFVEVQEQPQTLVELMIHRTGGSLGQVAVEWRVVGGTATEG 207
 D GVF +YS D + + + +++ I R G+ G VAV R+

5 Sbjct: 420 DPHGVFALYS-----
 DRQSIILIGQNLIRSIQINITRLAGTFGDVAVGLRISSDHKEQP 472

Query: 208 LDFIGAGEILTFAEGETKKTIVILTXXXXXXXXXXXXXXXXXXLVYTE-GGSRI-
 -LPSS-DT 263

10 + A L +G T K ++ LV G R
 +P+

Sbjct: 473
 IVTENAERQLVVKDGATYKVDVVPPIKNQVFLSLGSNFTLQLVTVMVLVGGFRFYGMPTILQE 532

15 Query: 264 VRVNIL-ANDNVAGI-VSFQTASRSVIGHEGEILQFHVIRTFPGR-
 GNVTVNWKI-IGQN 319

+ +L ++ A V F++ + ++ HV+ + G G ++V
 W

Sbjct: 533 AKSAVLVPSEKAANSQVGFESTAFQLMNITAGTS--
 20 HVMISRRGTYGALSVAWTTGYAPG 590

Query: 320 LEL-----
 NFANFSGQLFFPEGSLNTTLFVHLLDDNIPEEKEVYQVILYDVRTQGVPP 372

25 LE+ N G L F G +F+ P E + + L
 V++ P

Sbjct: 591 LEIPEFIVVGNMTPTLGSLSFSHGEQRKGVFLWTFPS--
 PGWPEAFVLHLSGVQSSA--P 646

Query: 373
 30 AGIALLDAAQGYAAVLTVASDEPHGVNLNLFALSSRFVLLQEANITIQLFINREFG-SLGAI 431
 G L G+ + A EP GV F+ SSR +++ E I+L + R

FG I

Sbjct: 647 GGAQL--RSGF-----
 IVAIEPMPGVFQFSTSSRNIIIVSEDTQMIRLHVQRLFGFHSDLI 699

35 Query: 432 NVTYTTVPGLS-LKN-QTV--GNLA----EPEVDF-
 VPIIGFLILEEGETAAAINITIL 482

V+Y T G L++ + V G L + EVDF + II L E E
 IN+T +

40 Sbjct: 700 KVSYQTTAGSAKPLEDFEPVQNGELFFQKFQTEVDFEITIINDQ-
 LSEIEEFFYINLTSV 758

Query: 483 E 483

45 Sbjct: 759 E 759

Score = 142 (21.3 bits), Expect = 5.6e-05, Sum P(3) = 5.6e-05
 Identities = 54/175 (30%), Positives = 76/175 (43%)

50 Query: 1435

DLPELEEGFIVTITEVNLVNSDFSTGQPSVRRPGMEIAEIMIEENDDPRGIFMFHVTRGA 1494
 DL + G+ TI E N + D QP + I I+I +ND+ GI

F

Sbjct: 16 DLYDFGRGYDFTIQE-NGLQID----QPP-
 55 EIGNISIVRIIMKNDNAEGIIIEFDPK--- 66

Query: 1495 GEVITAYXXXXXXXXXXXXXXXXXAGSFGAVNVYW--
 KASPDASAGLEDFKPSHGILEFADK 1552

TA+E

G++G V + ++S S G D+

+ F

Sbjct: b7 ---

YTAFEVEEDVGLIMIPVVRLHGTGYVTADFISQSSSASPGGVDYILHGSTVTFQHG 123

5

Query: 1553

QVTAMIEITIIDDAEFELTETFNISLISVAGGGRLGDDVVVTVVIPQNDSPFGVFGF 1609

Q + I I+IIDD E E E I L GG LG +V ++I

++DSPFGV F

10 Sbjct: 124

QNLSFINISIIDDNESEFEEPIELLTGATGGAVLGRHLVSRIIIAKSDSPFGVIRF 180

Score = 125 (18.8 bits), Expect = 4.0e-25, Sum P(3) = 4.0e-25

Identities = 77/308 (25%), Positives = 134/308 (43%)

15

Query: 1141 LVGAHSHIYELAYISSHS-----DFIP-

SSGELIFEPGEREATIAVNILDDTVPEKEES 1193

L G HS + +++Y ++

DF P +GEL F+ + E + I++D

+ E EE

20 Sbjct: 691

LFGFHSDLIKVSYQTTAGSAKPLEDFEPVQNGELFFQKFQTEVDFEITIINDQLSEIEEF 750

Query: 1194 FKVQLKNP--

KGGAEIGINDSVTITILSNDDAYGIVAFQAQNSLYKQVEEMEQLSLVTLNV 1251

F + L + +G + +N S + + D + ++ N

25

D L +++

Sbjct: 751 FYINLTSVEIRGLQKFDVNWSPRLNL---DFSVAVITILDN-----

DDLGMADI 796

30 Query: 1252

ERLKGTYGRITIAWEADGSISDIFPTSGVILFTEGQVLSTITLTLADNIPELSEVVIVT 1311

++ T+A D ++ + S L T + + + T + + E

+ V +

Sbjct: 797 -----SFPETTVAVAVDITLIPVETESTTYLSTS-

35 KTTTILQPTNVVAIVTEATGVSAIP 850

Query: 1312

LTRITTEGVEDSYKGATIDQDRSKSVITTLPNDSFGVLGWRRAASVFIRVAEPKENT-TT 1370

+T G T V T N S G + V+I E

40 K T T

Sbjct: 851 EKLVTLHG-----TPAVSEKPDVATVTANVSIHGTFSLGPSIVYIE-

EEMKNGTFNT 901

Query: 1371 LQLQIARDKGLLGDIHLRA-----QPNFL-----LHVDNQ--

45 ATENEDYVLQETI 1415

++ I R G G+++I ++

+PN L + N A E

ED+ Q

Sbjct: 902

AEVLIRRTGGFTGNVSITVKTFGERCAQMEPNALPFRGIYGISNLTWAVEEEDFEEQTLT 961

50

Query: 1416 IIMKENIKEAHA EVSILPDDLPELEEGFIVTIT 1448

+I + +E V IL DD PE +E F V +T

Sbjct: 962 LIFLDGERERKVSQILDDDEPEGQEFFYVFLT 994

55 Score = 123 (18.5 bits), Expect = 6.0e-28, Sum P(3) = 6.0e-28

Identities = 91/372 (24%), Positives = 150/372 (40%)

Query: 386 VLTVEASDEPHGVNLFALSSRFVLLQEA--NITI---
QLFINREFGSLGAINVTYTTV-- 438

V TV A+ HG F+L V ++E N T ++ I R G G
+++T T

5 Sbjct: 868 VATVTANVSIHGT--
FSLGPSIVYIEEEMKNGTFNTAEVLIRRTGGFTGNVSITVKTFGE 925

Query: 439 -----PGMLSLKN-QTVGNL--
AEPEVDFVPIIGFLILEEGETAAAINITILEDVPEL 489

10 P L + + NL A E DF LI +GE +++
IL+DD PE

Sbjct: 926
RCAQMEPNALPFRGIYGISNLTWAVEEEDFEEQTLTLIFLDGERERKVSQILDDDEPEG 985

15 Query: 490 EEFVLVNLTYVGLTMAASTSFPPRLDSEGLTA--QVIIDANDGARGVI---
EWQQRSEFV 544

+E+F V LT D G A VII +D G+I E
QS E+

Sbjct: 986 QEFFYVFLT----
20 NPQGGAQIVEGKDDTGFAAFAMVIITGSDLHNGIIGFSEESQSGLEL 1041

Query: 545 NE--THGSLTLVAQRS-REPLGHVSLF--
VYAQNLEAQVGLDYIFTMILHFADGERYKN 599

E L L+ R V +F V + D + L
25 G

Sbjct: 1042
REGAVMRRHLIVTRQPNRAFEDVKVFWRVTLNKTVVVLQKDGVLNMEELQSVSGTTTCT 1101

Query: 600 -----VNIMILDDDIPEGDEKFQILITNPSPGLELGKNT-
30 IALIIIVLANDDGPGLVLSF 651

++I + + +P+ + F + L + G + + A I +L
+D+ ++ F

Sbjct: 1102
MGQTKCFISIELKPEKVPQVEVYFFVELYEATAGAAINNSARFAQIKILESDESQSLVYF 1161

35 Query: 652 NNSEHFFLREPTALYVQESVAVLYIVREPAQGLFGTVTVQFIVTEVNSSNE-
-SKDLTPS 709

+ + A + L + R+ GL ++V F E+ S+
++P+

40 Sbjct: 1162 SVGSRLAVAHKKATLIS-----LQVARDSGTGLM--
MSVNFSTQELRSAETIGRTIISPA 1214

Query: 710 ---KGYIVLEEGVRFKALQISAILD 731
K +++ E + F+ Q S +LD

45 Sbjct: 1215 ISGKDFVITEGTLVFEPGQRSTVLD 1239

Score = 120 (18.0 bits), Expect = 1.8e-22, Sum P(3) = 1.8e-22
Identities = 77/316 (24%), Positives = 127/316 (40%)

50 Query: 1255 KGTYGRITIAWE---ADGS-----
ISDIFPTSGVILFTEGQVLSTITLILADNIPEL 1304

+GTYG +++AW A G + ++ PT G + F+ G+ + L
P

Sbjct: 573
55 RGTYGALSAWTTGYAPGLEIPEFIVVGNMPTLGSLSFSHGQKGVFLWTFPS--PGW 630

Query: 1305
SEVVIVTLTRITTEGVEDSYKGATIDQDRSKSVITTLPNDSPFGLVGWRAASVFIRVAEP 1364

E ++ L+ GV+ S G Q RS ++ + P G+ + +S

I V+E

Sbjct: 631 PEAFLVHLHS-----GVQSSAPGGA--QLRSGFIVAEI---
EPMGVFQFSTSSRNIIIVSE- 679

5

Query: 1365 KENTTTTLQLQIARDKGLLGDIHLRAQPNFLLHVDNQTENEDYV-
LQETIIIMKENIK 1423

+T ++L + R G D+ I + Q A ED+ +Q

+ ++

10

Sbjct: 680 --DTQMIRLHVQRLFGFHS DL-IKVS YQTTA-----
GSAKPLEDFEPVQNGELFFQKFQT 731Query: 1424 EAHAEVSILPDDLPELEEGFIVTITEVNLVN-
SDFSTGQPSVRRPGMEIAEIMIEENDDP 1482

15

E E++I+ D L E+EE F + +T V + F +A I

I +NDD

Sbjct: 732

EVDFEITIINDQLSEIEEFFYINLTSVEIRGLQKFDVNWSPRLNLD FSVAVITILDND DL 791

20

Query: 1483 RGI-FMFHVTRGAGEVITAY---
EXXXXXXXXXXXXXXXXXXAGSFGAVNVYWKASPD SAGLE 1538

G+ F T A V T E V +

+A+ SA E

Sbjct: 792

25

AGMDISFPETTVA VAVD TT LIPVETESTTYLSTSKTTTILQPTNVVAIVTEATGVSAIPE 851

Query: 1539 DFKPSHGILEFADKQVTAMIEITIIDD AEFEL 1570

HG ++K A + + F L

Sbjct: 852 KLVTLHGTPAVSEKPDVATVTANVSIHGTFSL 883

30

Score = 113 (17.0 bits), Expect = 9.4e-34, Sum P(3) = 9.4e-34
Identities = 28/87 (32%), Positives = 50/87 (57%)Query: 1156 SHSDFIPSSGELIFEPGEREATIAVNILDDT--
VPEKEESFKVQLKNPKGGAEIG-INDS 1212

35

S DF+ + G L+FEPG+R + V + +T + + F++ L +PKGGA

I + +

Sbjct: 1216

SGKDFVITEGTLVFEPGQRSTVLDVILTPETGSLNSFPKR FQIVLFD PKGGARIDKVYGT 1275

40

Query: 1213 VTITILSNDDAYGIVAF AQNSLYKQVEE 1240

IT++S+ D+ I A + L++ V +

Sbjct: 1276 ANITLVSDADSQAIWGLA-DQLHQPVND 1302

45

Score = 93 (14.0 bits), Expect = 4.1e-32, Sum P(3) = 4.1e-32
Identities = 57/222 (25%), Positives = 90/222 (40%)Query: 1404 TENEDYVL--QETIIIMKENIKEAHAE---VSILPDDLPEL-----
EEGFIVTITEVN 1451

50

TE+ Y+ + T I+ N+ E VS +P+ L L E+

+ T+T

Sbjct: 816

TESTTYLSTSKTTTILQPTNVVAIVTEATGVSAIPEKLVTLHGTPAVSEKPDVATVTANV 875

55

Query: 1452 LVNSDFSTGQPSVRRPGMEIAEIMIEENDDPRGIFMFHVTRGAGEV-
ITAYXXXXXXXXX 1510

++ FS G PS+ + I E M + + + G V IT

Sbjct: 876 SIHGTFSLG-PSI----
 VYIEEEMKNGTFNTAEVLIRRTGGFTGNVSITVKTGGERCAQM 930

Query: 1511

5 XXXXXXXXAGSFGAVNVYWKASPD SAGLED FKPSHGILEFADKQVTAMIEITIIDDAEFEL 1570
 G +G N+ W EDF+ L F D + + +

I+DD E E

Sbjct: 931 EPNALPFRGIYGISNLTWAVEE-----
 EDFEEQTLTLIFLDGERERKVSQILDDDEPEG 985

10

Query: 1571 TETFNISLISVAGGGRL--GDD-----VVVTVVIPQNDSPFGVFGFEEKTVS
 1615

E F + L + GG ++ G D V+I +D G+ GF E++ S
 Sbjct: 986 QEFFYVFLTNPQGGQIVEGKDDTGFAAFAMVIITGSDLHNGIIGFSEESQS
 15 1037

Score = 93 (14.0 bits), Expect = 1.0e-18, Sum P(3) = 1.0e-18
 Identities = 51/238 (21%), Positives = 107/238 (44%)

20 Query: 600 VNIMILDDDIPEGDEKFQILITNPSPGLELGKNT-
 IALIIVLANDDGPVLSFNNSEHFF 658

++I + + +P+ + F + L + G + + A I +L +D+ ++
 F+

Sbjct: 1109
 25 ISIELKPEKVPQVEVYFFVELYEATAGAAINNSARFAQIKILESDESQSLVYFSVGSRLA 1168

Query: 659 LREPTALYVQESVAVLYIVREPAQGLFGTVTVQFIVTEVNSSNE--
 SKDLTPS---KGYI 713

++P+ K ++ + A + L + R+ GL ++V F E+ S+
 30

Sbjct: 1169 VAHKKATLIS-----LQVARDSGTGML--
 MSVNFSTQELRSAETIGRTIISPAISGKDFV 1221

Query: 714 VLEEGVRFKALQISAILDT--EPE---MDEY---FVCTLFNPTGGARLG-
 35 VHVQTLITVL 764

+ E + F+ Q S +LD PE ++ + F LF+P GGAR+ V+
 IT++

Sbjct: 1222
 40 ITEGTLVFEPGQSTVLDVILTPETGSLNSFPKRFQIVLFDPKGGARIDKVYGTANITLV 1281

Query: 765 QNQAPLGLFSISAVENRATSIDI-
 EEANRTVYLVNSRTNGIDLAVSVQWETVSETAFGMR 823

+ ++ ++ ++ + DI T+ + V+ T D +S +
 45 Sbjct: 1282 SDADSQAIWGLADQLHQPVNDILNRVLHTISMKVA-
 TENTDEQLSAMMHLIEKIT--TE 1338

Query: 824 GMDVVFSV 831
 G FSV

50 Sbjct: 1339 GKIQAFSV 1346

Score = 92 (13.8 bits), Expect = 9.5e-25, Sum P(3) = 9.5e-25
 Identities = 44/177 (24%), Positives = 82/177 (46%)

55 Query: 680

PAQGLFGTVTVQFIVTEVNSSNESKDLTPSKGYIVLEEGVRFKALQISAILDTEPEMDEY 739
 P +G++G + + V E + E + LT ++ +G R + + + + D
 EPE E+

Sbjct: 936 PFRGIYGISNLTWAVEEEDF--EEQTLT-----
LIFLDGERERKVSQILDDDEPEGQEF 988

Query: 740 FVCTLFNPTGGARL-----

5 GVHVQTLITVLQNAQAPLGLFSISAVENRATSIDIEEAN- 791
F L NP GGA++ G ++ + + G+ S E +

+++ E

Sbjct: 989 FYVFLTNPQGGGAQIVEGKDDTGFAAFAMVIITGSDLHNGIIGFS--
EESQSGLELREGAV 1046

10 Query: 792 -RTVYLVNSRT-NGIDLAVSVQWE-TVSETAF-----

GMRGMDVVFSVFQSFLEDESASGW 843

R ++L V+R N V V W T+++T G+ M+ + SV +

Sbjct: 1047

15 MRRRLHLIVTRQPNRAFEDVKVFWRVTLNKTVVVLQKDGVLNMEELQSVSGTTTCTMGQTK 1106

Query: 844 CFFTLE 849

CF ++E

Sbjct: 1107 CFISIE 1112

20 Score = 91 (13.7 bits), Expect = 6.6e-32, Sum P(3) = 6.6e-32
Identities = 49/153 (32%), Positives = 70/153 (45%)

Query: 1466

25 RPGMEIAEIMIEENDDPRGIFMFHVTRGAGEVITAYXXXXXXXXXXXXXXXXXAGSFGAVN 1525
R G +AEI +P G+F F + + +I + +

+
Sbjct: 652 RSGFIVAEI-----EPMGVFQFSTS--
SRNIIVSEDTQMIRLHVQRLFGFHSD---LIK 700

30 Query: 1526 VYWKASPDASAG-LEDFKP-

SHGILEFADKQVTAMIEITIIDDAEFELTETFNISLISVAG 1583

V ++ + SA LEDF+P +G L F Q EITII+D E+ E F

I+L SV

35 Sbjct: 701

VSYQTTAGSAKPLEDFEPVQNGELFFQKFQTEVDFEITIIINDQLSEIEEFFYINLTSVEI 760

Query: 1584 GG-----RLGDDVVVTVV-IPQNDSPFGV-FGFEETVS 1615

G RL D V V+ I ND G+ F E TV+

40 Sbjct: 761 RGLQKFDVNWSRPNLDFSVAVITILDNDLAGMDISFPETTVA 804

Score = 65 (9.8 bits), Expect = 8.8e-29, Sum P(3) = 8.8e-29
Identities = 26/99 (26%), Positives = 50/99 (50%)

45 Query: 1232 NSLYKQVEEMEQDSLVTNLNVERLKGTYGRITIAWEADGS----ISDIF--
PTSGVILFTE 1285

NS K+ + + D +++++ GT IT+ +AD ++D P

+ IL

Sbjct: 1250 NSFPKRFQIVLFDPKGGARIDKVYGT-

50 ANITLVSDADSQAIWGLADQLHQPVNDDIL--- 1305

Query: 1286 GQVLSTITLTILADNIPELSEVVIVTLTRITTEGVEDSYKGAT 1328

+VL TI++ + +N E ++ + +ITTEG ++ A+

Sbjct: 1306 NRVLHTISMKVATENTDEQLSAMMHLIEKITTEGKIQAFSVAS 1348

55 Score = 48 (7.2 bits), Expect = 1.9e-27, Sum P(3) = 1.9e-27
Identities = 23/115 (20%), Positives = 44/115 (38%)

Query: 1499 TAYXXXXXXXXXXXXXXXXXAGSFGAVNVYWKAS-----
 PDSAGLEDFKPSHGILEFAD 1551

TA++

G++GA++V W

P+ + + P+

G L F+

5 Sbjct: 554

TAFQLMNITAGTSHVMISRRGTYGALSVAWTTGYAPGLEIPEFIVVGNMTPTLGSLSFH 613

Query: 1552 KQVTAMIEITIIDDAEFELTETFNISLI--

SVAGGGRLGDDVVVTVVIPQNDSPFGVFGF 1609

10 + + + ++S + S GG +L +V +

P GVF F

Sbjct: 614 GEQRKGVFLWTFPSPGWPEAFVLHLSGVQSSAPGGAQLRSGFIVAEI-----

EPMGVFQF 668

15

Pedant information for DKFZphamy2_10p7, frame 3

Report for DKFZphamy2_10p7.3

20

[LENGTH] 1615

[MW] 177600.58

[pI] 4.37

25 [HOMOL] TREMBL:AF055084_1 gene: "VLGR1"; product: "very
 large G-protein coupled receptor-1"; Homo sapiens very large G-
 protein coupled receptor-1 (VLGR1) mRNA, complete cds. 5e-24

[BLOCKS] BP01493A

[BLOCKS] BLO0713B Sodium:dicarboxylate symporter family proteins

30

[BLOCKS] PRO1003A

[BLOCKS] PRO0412C

[BLOCKS] BLO0824E

[PIRKW] heart 1e-08

35

[PIRKW] ion transport 1e-08

[PIRKW] transmembrane protein 3e-08

[PIRKW] phosphoprotein 2e-08

[PIRKW] membrane protein 1e-08

[PROSITE] MULTICOPPER_OXIDASE1 1

40

[KW] All_Beta

[KW] LOW_COMPLEXITY 2.60 %

45 SEQ DAWADAWALYTCATLCLKEQACSAFSFFSASEGPQCFWMTSWISPAVNNSDFWYRK NMT

SEGxxxxxxxxxxxxx.....

PRD ccchhhhhhhhhchhhhhhhhhheeeeeecccccceeeeeecccccceeeeeecceee

50 SEQ RVASLFGQAVAGSDYEPVTRQWAIMQEGDEFANLTVSILPDDFPENDESFLISLLEVHL

SEG

PRD eeeeeccccccccccccceeeeeecccccceeeeeecccccccchhhhhhhhhhhhhh

SEQ MNISASLKNQPTIGQPNISTVVIALNGDAFGVFVIYSISPNTSEDGLFVEVQEQPQTLVE

SEG

PRD hccccccccccccccccceeeeeecccccceeeeeecccccccccceeeeeecceee

55

SEQ LMIHRTGGSLGQVAVEWVVGGTATEGLDFIGAGEILTFAEGETKKTIVILTILDDSEPED

SEGxxxxxxxxxx

PRD eeeeeccccccccceeeeeecccccccccccccceeeeeecccccceeeeeecceee

-49-

```

SEG .....
PRD eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee

5  SEQ LVGAHSHIYELAYISSHSDIFPSSGELIFEPGEREATIAVNILDDTVPEKEESFKVQLKN
   SEG .....
   PRD eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeec

10  SEQ PKGGAEIGINDSVTITILSNDDAYGIVAFQAQNSLYKQVEEMEQDSLVTLNVERLKGTYGR
   SEG .....
   PRD cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

15  SEQ ITIAWEADGSISDIFPTSGVILFTEGQVLSTITLTILADNIPELSEVVIVTLTRITTEGV
   SEG .....
   PRD eeeeeeeeeccccccccccccccccccccccccccccccccccccccccccccccccccccc

20  SEQ EDSYKGATIDQDRSKSVITTLPNDSFGVLGVRAASVFIRVAEPKENTTTLQLQIARDKG
   SEG .....
   PRD cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

25  SEQ LLDIAIHLRAQPNFLLHVDNQATENEDYVLQETIIIMKENIKEAHAESVILPDDLPELE
   SEG .....
   PRD cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

30  SEQ EGFIVTITEVNLVNSDFSTGQPSVRRPGMEIAEIMIEENDPRGIFMFHVTRGAGEVITA
   SEG .....
   PRD cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

35  SEQ YEVPPLNLVLQVPVVRLAGSFGAVNVYWKASPDAGLEDFKPSHGILEFADKQVTAMIEI
   SEG ..xxxxxxxxxxxxxxxx.....
   PRD eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeecce

   SEQ TIIDDAEFELTETFNISLISVAGGRLGDDVVVTVVVIPQNDSPFGVFGFEKTVS
   SEG .....
   PRD eeechhhhhhhhhccccccccccccccccccccccccccccccccccccccccccc

```

Prosites for DKFZphamy2_10p7-3

40 PS00079 151->172 MULTICOPPER_OXIDASE1 PD0C00076

(No Pfam data available for DKFZphamy2_10p7-3)

DKFZphamy2_11d2

5 group: transmembrane protein

DKFZphamy2_11d2 encodes a novel 552 amino acid protein without similarity to known proteins.

10 The novel protein contains 2 transmembrane regions.
No informative Blast results; no predictive prosite, pfam or scope motife.

15 The new protein can find application in studying the expression profile of amygdala-specific genes and as a new marker for amygdala cells.

unknown protein

20

Pedant: TRANSMEMBRANE 2

Sequenced by EMBL

25 Locus: /map="11bp13.3"

Insert length: 2939 bp

Poly A stretch at pos. 2920, polyadenylation signal at pos. 2869

30

1 GGC GGG GTG AGG CCG CGG GCG CAG GTT CCA CCT GGG GCT TG CGA AGG CACA
51 GAT TCCCC GT CCAC AGCT CA CGAC CAG ATG CACC AGC AGG AGT CCA CAT C
101 GAG GAC GT CC TCC GGG CACT CCC ACG ACCA GTG ACC AGG A GTT AAA CTTT
151 GGG ATG TG CC CGT GAT GTT G GACC ACA AGG ACT TAG AGG C CGA AAT CCA C
35 201 CCCT TGAAAA ATGA AGAA AG AAA ATCG CAG GAAA ATCT GG GAA ATCC ATC
251 AAAAA ATG AG GATA AC GT GA AAAG CG CGCC TCC ACAG TCC CGG CTCT CCC
301 GGT GCC GAG C GGC GGC GTT TTT CTTT CAT TGT TTT CTCT G CCT TTTT GTG
351 GTG TTC GT CG TCT CAT TCGT CAT CCC GTGT CCAG ACC GGC CGG CGT CACA
401 GCG AAT GT GG AGG ATAG ACT ACAG TGCC GC TGT TAT CTAT GACT TTT CTGG
40 451 CTG TGG AT GA TATA AAC GGG GAC AGG ATCC AAG ATG TTCT TTTT CTTT AT
501 AAA AAC ACCA ACAG CAG CAA CAAT TTT CAG C GAT CCT GTG TGG AC GA AGG
551 CTTT TCCT CT CCCT GCAC CT TTGC AGCT GC TGT GTC GGGG GCC AAC GGC A
601 GCAC GCT CTG GGAG AGAC CT GTGG CCA AG ACGT GGC CT CGT GGAG TGT
651 GCT GTG CCCC AGCCA AGAG G CAG TGAG GCA CCTT CTGC CT GCAT CCT GGT
45 701 GGG CAG ACC AGTT CTTT CA TTGC AGT CAA CTT GTT CACA GGG GAA ACCC
751 TGT GGA ACCA CAG CAG CAG C TTC AGC GGG A ATG CGT CCA T CCT GAG CCT
801 CTG CTG CAG G TGC CTG ATGT GGC GGC GAT GGG GCCCC AG ACCT GCT GGT
851 TCT CAC CCAG GAG CGG GAG G AGG TTAG TGG CCAC CTCT AC TCC GGC AGCA
901 CCG GGC ACCA GATT GGC CTC AGAG GCAG CC TTGG TGT GGA CGG GGA AGT
50 951 GGCT TCCT CC TTCAC GT CAC CAGG ACAG GT GCCC ACTACA TCCT CTTT CC
1001 CTG CGC AAG C TCC CTCT GCG GCT GCT CTGT GAAG GTCT C TAC GAG AAG G
1051 TGAC CGGG AG CGG CGG CCG TTCA AGAG TG ACC CGC ACTG GGAG AGCAT G
1101 CTCA ATGCCA CCAC CCAG C GAT GCT TTT CC CAC AGCT CTG GAG CAG TGCG
1151 CTAC CTG ATG CAT GT CCCC G GGA AC GCG G TGC AGAT GTG CTT CTGT GG
55 1201 GCT CAG AGG CTT CGT GCT G CTGG AC GGC AGG AGCT GAC CCCT CGT GG
1251 ACAC CCA AGG CAG CCC ATGT CCT GAG AAAA CCC ATCT TC G GCG CTA CAA
1301 ACC AGA CAC TTGG CTGT AG CCG TTG AAAA CGGA ACT GGC ACC GAC AGAC
1351 AGAT CCT GTT TCT GGAC CT GGC ACT GGAG CCGT CCT GTG TAG CCT AGCC

1401 CTCCCGAGCC TCCCTG6GGG TCCACTGTCC GCCAGCCTGC CGACCGCAGA
 1451 CCACCGCTCA GCCTTCTTCT TCTGGGGCCT CCACGAGCTG GGGAGCACCA
 1501 GCGAGACGGA GACCGGGGAG GCCCGGCACA GCCTGTACAT GTTCCACCCC
 1551 ACCCTGCCGC GCGTGCTGCT GGAGCTGGCC AATGTCTCTA CCCACATTGT
 5 1601 CGCCTTTGAC GCCGTCTGT TTTAGCCAAG CCGCCACGCC CCTACATCC
 1651 TTCTGACAGG CCCGGCAGAC TCAGAGGCAC CCGGCCTGGT CTCTGTGATC
 1701 AAGCACAAGG TGGGGGACCT TGTCCCAAGC AGCAGGGTGG TCCGCCTGGG
 1751 TGAGGGTGGG CCAGACAGTG ACCAAGCCAT CAGGGACCGG TTCTCCCGGC
 1801 TGGGTACCA GAGTGAGGCG TAGAGGCACG CCAGCCAGAG CCTGTGGAGA
 10 1851 GACTCCGCCT GCTGACACTA AACGTCCTGG GAAGTGGGCC CTTCCTGGG
 1901 TCTCTGCACT GACTCCCCCA CTCCTGACCC TGGTGATGGT CGCCACTGGG
 1951 CAGCAGCAGC CTTACCAGTC CTCCATGATC ACACCCAGGG ACCTGCATGG
 2001 GTGAGGGGAG ACCCTGGGCG TCTCTCCCGC CCAGCATCCT CCTGAGTCC
 2051 CCACACAGGG CCTCACTCTG CACCCACCA GGGTCCCGCT CACACCAGGC
 15 2101 AGCCTTCATA GTGGTCTCCC TGGCCACCTT GGGCAGAGCT GGGTCATGCA
 2151 GCACCCCATC CTTACCCGGT GCCCTCTCCT TGCCAGCTTC TCCCCAGGCC
 2201 AGAGCGGCCA TCGCGTAGAA AGAACCAGGG TGTCCCGGGG ACAGGCCGTC
 2251 CCCCACCCCA TCCTGTAGAG TCCATTCCCC TTTTCCCTCC TGTGCTCTGT
 2301 CCCCCAAGGA GTCATGGAAC TCAGGGTACT GGGCCTCAAC GGGAACCTGA
 20 2351 GACAGCTTCC AGCTTCGCAG CCTTCCCGG AGCTACAGGG GGATCCTCTA
 2401 GCATGGGGGG TGTGACTTGG TTCTTTGAC CAGGTCTGT GAGGAAGCCT
 2451 GGAGCAAGGG TCTCCCCAG CAGGATGGGT GGGGCTGCT CTGGAGCTGA
 2501 GCGCGTGGCC GCTCACAGGT GTCCTTAGTG GTGTTGCAGC TGTCTACTGG
 2551 CTGCTGTGTC TGTGAATATC CCAAGGAACT GGCTGTGGAA TGCCTGTTTG
 25 2601 GGTCACTCTG TGCCCTCTCA GTAGACACTG GAGCTGCTCT GTCCCTGAAG
 2651 AGGCCCCGTG CCCCAGGCAT GGCAAGCGCC TGCCTCTCCC CTTCGGGTGC
 2701 TCACACGCCC ACGCCGTGCC ACCCGATGCA GGAATCACCT CTGTGCCCTG
 2751 CTGCTCCTGA GGCCCAAGGG CAGCCATGGT GCTCTGTACT GCTCGGGCCG
 2801 CCCAGGTCAC AGAGCCTGAG CTTCTGTAGC AAAGCAGCCT GATGACCCAC
 30 2851 CCACCAAGGA AGAAAGCAGA ATAAACATT TGGCACTGCC TGAAAAACCC
 2901 CGGTGGTCAG GCGTGAGCCT AAAAAAAAAA AAAAAAAAAA

BLAST Results

35

No BLAST result

40

Medline entries

No Medline entry

45

Peptide information for frame 2

50

ORF from 2555 bp to 2839 bp; peptide length: 95
 Category: questionable ORF
 Classification: unclassified

55

1 MCCEYPKELA VECVFGSVCA LSVDTGAALS LKRPRAPGMA SACLSPSGAH
 51 TPTPCHPMQD SPLCLAAPEA QGQPWCSVLL GPPRSQSLSF VAKAA

BLASTP hits

No BLASTP hits available

5 Alert BLASTP hits for DKFZphamy2_11d2, frame 2

TREMBL:MMIGCF_2 Mouse ig gamma2a-b(c57bl/6 allele) c gene and
secreted
tail. N = 1, Score = 73, P = 0.1

10

>TREMBL:MMIGCF_2 Mouse ig gamma2a-b(c57bl/6 allele) c gene and
secreted
tail.

15

Length = 334

HSPs:

Score = 73 (11.0 bits), Expect = 1.1e-01, P = 1.0e-01
Identities = 16/49 (32%), Positives = 27/49 (55%)

20

Query: 44 LSPSGAHTPTPCHPMQDSPLCLAAPEAQGQPWCSVLLGPPRSQSLSFVA 92
+ P T PC P+++ P C AAP+ G P SV + PP+ + + ++
Sbjct: 96 IEPRVPITQNPCPLKECPPC-AAPDLLGGP--SVFIFPPKIKDVLMS
141

25

Peptide information for frame 3

30

ORF from 165 bp to 1820 bp; peptide length: 552
Category: putative protein
Classification: Transmembrane proteins unclassified

35

1 MLDHKDLEAE IHPLKNEERK SQENLGNPSK NEDNVKSAPP QSRLSRCRAA
51 AFFLSLFLCL FVVFVVSFVI PCPDRPASQR MWRIDYSAAV IYDFLAVDDI
101 NGDRIQDVLF LYKNTNSSNN FSRSCVDEGF SSPCTFAAAV SGANGSTLWE
151 RPVAQDVALV ECAVPQPRGS EAPSACILVG RPSSFIAVNL FTGETLWNHS
40 201 SSFSGNASIL SPLAQVPDVD GDGAPDLLVL TQEREEVSGH LYSGSTGHQI
251 GLRGS LGVDG ESGFLLHVTR TGAHYILFPC ASSLCGCSVK GLYEKVTGSG
301 GPFKSDPHWE SMLNATTRRM LSHSSGAVRY LMHVPGNAGA DVLLVGSEAF
351 VLLDGQELTP RWTPKAAHVL RKPIFGRYKP DTLAVAVENG TGTDRQILFL
401 DLGTGAVLCS LALPSLPGGP LSASLPTADH RSAFFFFWGLH ELGSTSETET
45 451 GEARHSLYMF HPTLPRVLE LANVSTHIVA FDAVLFEPSR HAAYILLTGP
501 ADSEAPGLVS VIKHKVRDLV PSSRVVRLGE GGPDSDAQAIR DRFSRLRYQS
551 EA

50

BLASTP hits

No BLASTP hits available

55 Alert BLASTP hits for DKFZphamy2_11d2, frame 3

No Alert BLASTP hits found

Pedant information for DKFZphamy2_11d2, frame 2

Report for DKFZphamy2_11d2.2

5

[LENGTH] 95
 [MW] 9757.38
 [pI] 6.68
 10 [BLOCKS] PRO0521E
 [KW] Alpha_Beta

15 SEQ MCCEYPKELAVECVFGSVCALSVDTGAALSLKRPRAPGMASACLSPSGAHTPTPCHPMQD
 PRD cccchhhhhhhhhccceeeeeeecccchhhhhcccccccccccccccccccccccccccccc

SEQ SPLCLAAPEAQGQPWCSVLLGPPRSQSLSFVAKAA
 PRD cccccccccccccceeecccccccchhhhhccc

20

(No Prosite data available for DKFZphamy2_11d2.2)

(No Pfam data available for DKFZphamy2_11d2.2)

25

Pedant information for DKFZphamy2_11d2, frame 3

Report for DKFZphamy2_11d2.3

30

[LENGTH] 552
 [MW] 59659.68
 [pI] 5.84
 35 [BLOCKS] PRO0211G
 [BLOCKS] BL00288C Tissue inhibitors of metalloproteinases
 proteins
 [BLOCKS] PRO0436A
 [KW] TRANSMEMBRANE 2
 40 [KW] LOW_COMPLEXITY 8.15 %

45 SEQ MLDHKDLEAEIHPLKNEERKSQENLGNPSKNEEDNVKSAPPQSRLSRCRAAAFFLSLFLCL
 SEGxxxxxxxxx
 PRD cccchhhhhhhcc
 MEMMMMMMMM

50 SEQ FVVVFVVSFVIPCPDRPASQRMWRIDYSAAVIYDFLAVDDINGDRIQDVLFLYKNTNSSNN
 SEG xxxxxxxxxxxx.....
 PRD hhhhhhhcc
 MEM MMMMMMMMMM.....

55 SEQ FSRSCVDEGFSSPCTFAAAVSGANGSTLWERPVAQDVALVECAVPQPRGSEAPSACILVG
 SEG
 PRD ccc
 MEM

SEQ RPSSSFIAVNLFTGETLWNHSSSFSGNASILSPLLQVPDVGDDGAPDLLVLTQEREEVSGH

-55-

DKFZphamy2_11n4

5 group: nucleic acid management

DKFZphamy2_11n4 encodes a novel 1091 amino acid protein with similarity to RAD18 of *Schizosaccharomyces pombe* and YLR383w of *Saccharomyces cerevisiae*.

10 The novel protein contains a ATP/GTP-binding site motif A (P-loop). It has similarity to RAD18 acts in a DNA repair pathway for removal of UV-induced DNA damage. YLR383w of *Saccharomyces cerevisiae* is a recombination repair protein.

15 The new protein can find application in modulation of DNA-repair and as a new tool for manipulation of nucleic acids.

similarity to RAD18 (*Schizosaccharomyces pombe*)

20 comment on P53692:
FUNCTION: ACTS IN A DNA REPAIR PATHWAY FOR REMOVAL OF UV-INDUCED DNA DAMAGE THAT IS DISTINCT FROM CLASSICAL NUCLEOTIDE EXCISION REPAIR AND IN REPAIR OF IONIZING RADIATION DAMAGE.

25 Sequenced by EMBL

Locus: /map="2"

30 Insert length: 3679 bp
Poly A stretch at pos. 3646, polyadenylation signal at pos. 3620

```

1  ACCGCGGTGG GCGCCGGGGC TCCCGGGAAT CTACCTTCTC CTGCGGCCGG
35  51  CACGCGGTTC CCAGGGGGGCC AGCGGCGGTC AGCCGAGGTC GAGACGCCCC
    101  CAGGGTGGCC TTAGCGGCCG GTCGTACCAC GGCAGCCCCG CCGATCAGGT
    151  TCCTTTGGGA GACTTCGACT TGTGGCGGAA ATGAACCGGA GAAGAATCCC
    201  AATTGGGAAT TGCGGAAAAC AGGACTCTAG GGTAGAGAAA GGTGTAGAA
    251  CCAATAGGGT TTGAGACCTG ATGGCCAAAA GAAAGGAAGA AAATTTTTC
40  301  TCTCCTAAAA ATGCCAAAAG GCCAAGACAA GAAGAATTGG AGGATTTTGA
    351  TAAAGATGGT GACGAAGACG AATGTAAAGG TACTACTTTG ACTGCAGCAG
    401  AAGTTGGAAT AATTGAGAGT ATTCACCTAA AAAACTTCAT GTGTCATTCA
    451  ATGCTTGGAC CTTTTAAGTT TGGTTCTAAT GTCAACTTTG TTGTTGGCAA
    501  CAATGGAAGT GGAAGAGTG CAGTACTCAC AGCTCTCATA GTCGGTCTTG
45  551  GTGGAAGAGC AGTTGCTACT AATAGAGGAT CCTCTTTAAA AGGTTTTGTG
    601  AAAGATGGAC AGAACTCTGC AGATATCTCA ATAACATTGA GGAACAGAGG
    651  AGATGATGCC TTTAAAGCCA GTGTGTATGG TAACTCTATA CTTATACAGC
    701  AACACATCAG CATAGATGGA AGTCGATCTT ATAAACTTAA AAGTGCAACA
    751  GGCTCCGTGG TTTCCACGAG GAAAGAAGAG CTGATTGCAA TTCTTGATCA
50  801  TTTTAACATC CAGGTGGATA ATCCAGTTTC TGTTTTAACA CAAGAAATGA
    851  GCAAGCAGTT CTTACAGTCT AAAAATGAAG GAGACAAATA CAAATTCTTC
    901  ATGAAAGCAA CGCAACTTGA ACAGATGAAG GAAGATTATT CATACATTAT
    951  GGAACGAAA GAAAGAACAA AGGAGCAGAT ACATCAAGGA GAAGAGCGGC
1001  TTAAGTGAAT AAAGCGCCAG TGTGTAGAGA AAGAGGAACG TTTTCAAGGT
55 1051  ATTGCTGGTT TAAGTACAAT GAAGACTAAT TTAGAGTCCT TGAAACATGA
    1101  AATGGCTTGG GCAGTGGTCA ATGAAATTGA AAAACAATTG AATGCCATCA
    1151  GAGATAATAT CAAAATTGGA GAAGATCGTG CTGCTAGACT TGACAGGAAA
    1201  ATGGAAGAAC AGCAGGTGAG ACTTAATGAG GCAGAACAAA AGTACAAGGA

```

1251 TATTCAAGAC AAAC TAGAAA AGATTAGTGA AGAGACAAAT GCACGAGCAC
1301 CAGAATGTAT GGCATTGAAA GCAGATGTTG TTGCTAAGAA AAGGGCCTAT
1351 AATGAAGCTG AGGTTTTATA TAACCGATCC TTAAACGAAT ATAAAGCATT
1401 AAAGAAAGAT GATGAGCAGC TTTGTAAACG AATTGAAGAG CTGAAAAAAA
5 1451 GTACTGACCA ATCTTTGGAA CCTGAACGGT TGGAAAAGACA AAAAAAATA
1501 TCTTGGTTAA AAGAGAGAGT AAAGGCCTTT CAAAATCAAG AAAATTTCAGT
1551 CAATCAAGAG ATCGAACAGT TTCAGCAAGC CATAGAAAAG GACAAAGAAG
1601 AACATGGCAA AATTAAGAGA GAAGAATTAG ATGTGAAGCA TGCACTGAGC
1651 TACAATCAGA GGCAACTGAA AGAATTGAAA GATAGTAAAA CTGATCGACT
10 1701 CAAAAGATTT GGCCCTAATG TTCCAGCTCT TCTTGAAGCC ATAGATGATG
1751 CTTATAGACA AGGCATTTT ACCTATAAAC CTGTAGGCCCT TTTAGGAGCT
1801 TGCATTGATC TTCGGGACCC AGAAGTTGCT TTGGCTATTG AATCTTGCTT
1851 AAAAGGGCTT CTGCAGGCCT ATTGTTGCCA TAATCATGCT GATGAAAGGG
1901 TCCTTCAGGC ACTCATGAAA AGGTTTTATT TACCAGGGAC CTCACGGCCA
15 1951 CCGATAATAG TTTCTGAGTT TCGGAATGAG ATATATGATG TAAGACACAG
2001 AGCTGCTTAT CATCCAGACT TTCCAACAGT TCTGACAGCT TTAGAAATAG
2051 ATAATGCGGT TGTGGCAAAT AGCCTAATTG ACATGAGAGG CATAGAGACA
2101 GTGCTACTAA TCAAAAATAA TTCTGTAGCT CGTGCAGTAA TGCAGTCCCA
2151 AAAGCCACCC AAAAAATTGA GAGAAGCTTT TACTGCTGAT GGTGATCAAG
20 2201 TTTTTCAGG ACGTTATTAT TCATCTGAAA ATACAAGACC TAAGTTCCCTA
2251 AGCAGAGATG TGGATTCTGA AATAAGTGAC TTGGAGAATG AGGTTGAAAA
2301 TAAGACGGCC CAGATATTAA ATCTTCAGCA ACATTTATCT GCCCTTGAAA
2351 AAGATATTAA ACACAATGAG GAACTTCTTA AAAGGTGCCA ACTACATTAT
2401 AAAGAACTAA AGATGAAAAT AAGAAAAAAT ATTTCTGAAA TTCGGGAACCT
25 2451 TGAGAACATA GAAGAACACC AGTCTGTAGA TATTGCAACT TTGGAAGATG
2501 AAGCTCAGGA AAATAAAAGC AAAATGAAAA TGGTTGAGGA ACATATGGAG
2551 CAACAAAAAG AAAATATGGA GCATCTTAAA AGTCTGAAAA TAGAAGCAGA
2601 AAATAAGTAT GATGCAATTA AATTCAAAAT TAATCAACTA TCGGAGCTAG
2651 CAGACCCACT TAAGGATGAA TTAAACCTTG CTGATTCTGA AGTGGATAAC
30 2701 CAAAAACGAG GGAAACGACA TTATGAAGAA AAACAAAAAG AACACTTGGA
2751 TACCTTAAAT AAAAAGAAAC GAGAACTGGA TATGAAAGAG AAAGAACTAG
2801 AGGAGAAAAAT GTCACAAGCA AGACAAATCT GCCCAGAGCG TATAGAAGTA
2851 GAAAAATCTG CATCAATTCT GGACAAAGAA ATTAATCGAT TAAGGCAGAA
2901 GATACAGGCA GAACATGCTA GTCATGGAGA TCGAGAGGAA ATAATGAGGC
35 2951 AGTACCAAGA AGCAAGAGAG ACCTATCTTG ATCTGGATAG TAAAGTGAGG
3001 ACTTTAAAAA AGTTTATTAA ATTACTGGGA GAAATCATGG AGCACAGATT
3051 CAAGACATAT CAACAATTTA GAAGGTGTTT GACTTTACGA TGCAAATTAT
3101 ACTTTGACAA CTTACTATCT CAGCGGGCCT ATTGTGGAAA AATGAATTTT
3151 GACCACAAGA ATGAAACTCT AAGTATATCA GTTCAGCCTG GAGAAGGAAA
40 3201 TAAAGCTGCT TTCAATGACA TGAGAGCCTT GTCTGGAGGT GAACGTTCTT
3251 TCTCCACAGT GTGTTTTATT CTTTCCCTGT GGTCCATCGC AGAATCTCCT
3301 TTCAGATGCC TGGATGAATT TGATGTCTAC ATGGATATGG TTAATAGGAG
3351 AATTGCCATG GACTTGATAC TGAAGATGGC AGATTCCCAG CGTTTTAGAC
3401 AGTTTATCTT GCTCACACCT CAAAGCATGA GTTCACTTCC ATCCAGTAAA
45 3451 CTGATAAGAA TTCTCCGAAT GTCTGATCCT GAAAGAGGAC AAACACATT
3501 GCCTTTTCAGA CCTGTGACTC AAGAAGAAAG TGATGACCAA AGGTGATTTG
3551 TAACTTAACA TGCCTTGTCC TGATGTTGAA GGATTTGTGA AGGGAAAAAA
3601 AATTCTGGAC TCTTTGATAT AATAAAATGA GACTGGAGGC ATTCTGAAAA
3651 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA

BLAST Results

55 No BLAST result

Medline entries

96069417:

Lehmann AR, Walicka M, Griffiths DJ, Murray JM, Watts FZ,
5 McCready S,Carr AM.; The rad18 gene of Schizosaccharomyces pombe defines a
new subgroup of the SMC superfamily involved in DNA
repair. Mol Cell Biol 1995 Dec;15(12):7067-80

10 99380167:

Mengiste T, Revenkova E, Bechtold N, Paszkowski J.; An SMC-like
proteinis required for efficient homologous
recombination in Arabidopsis. EMBO J 1999 Aug 16;18(16):4505-12

15

Peptide information for frame 1

20

ORF from 271 bp to 3543 bp; peptide length: 1091

Category: similarity to known protein

Classification: Nucleic acid management

25

Prosites motifs: RGD (126-128)

ATP_GTP_A (76-83)

30

1	MAKRKEENFS	SPKNAKRPRQ	EELEDFDKDG	DEDECKGTTL	TAAEVGIIES
51	IHLKNFMCHS	MLGPFKFGSN	VNFVVGNGS	GKSAVLTALI	VGLGGRAVAT
101	NRGSSLKGFV	KDGQNSADIS	ITLRNRGDDA	FKASVYGNSI	LIQQHISIDG
151	SRSYKLKSAT	GSVVSTRKEE	LIAILDHFNI	QVDNPVSVLT	QEMSKQFLQS
201	KNEGDKYKFF	MKATQLEQMK	EDYSYIMETK	ERTKEQIHQG	EERLTELRQ
251	CVEKEERFQS	IAGLSTMKTN	LESLKHEMAW	AVVNEIEKQL	NAIRDNIKIG
301	EDRAARLDRK	MEEQQVRLNE	AEQKYKDIQD	KLEKISEETN	ARAPECMALK
351	ADVVAKKRAY	NEAEVLYNRS	LNEYKALKKD	DEQLCKRIEE	LKKSTDQSLE
401	PERLERQKKI	SWLKERVKAF	QNEQNSVNQE	IEQFQQAIEK	DKEEHGKIKR
451	EELDVKHALS	YNQRQLKELK	DSKTDRLKRF	GNVPALLEA	IDDAYRQGHF
501	TYKPVGPLGA	CIHLRDPELA	LAIESCLKGL	LQAYCCHNHA	DERVLQALMK
551	RFYLPGTSRP	PIIVSEFRNE	IYDVRHRAAY	HPDFPTVLTA	LEIDNAVVAN
601	SLIDMRGIET	VLLIKNNSVA	RAVMQSQKPP	KNCREAFTAD	GDQVFAGRY
651	SENTRPKFL	SRDVSSEISD	LENEVENKTA	QILNLQQHLS	ALEKDIKHNE
701	ELLKRCQLHY	KELKMKIRKN	ISEIRELENI	EEHQSVDIAT	LEDEAQENKS
751	KMKMVEEHME	QKENMEHLK	SLKIEAENKY	DAIKFKINQL	SELADPLKDE
801	LNLADSEVDN	QKRGKRHYEE	KQKEHLDTLN	KKKRELDME	KELEEKMSQA
851	RQICPERIEV	EKSASILDK	INRLRQKIQ	EHASHGDREE	IMRQYQEAARE
901	TYLDLDSKVR	TLKKFIKLLG	EIMEHRFKTY	QQFRRCLTLR	CKLYFDNLLS
951	QRAYCGKMF	DHKNETLSIS	VQPGEGNKAA	FNDMRALSGG	ERSFSTVCFI
1001	LSLWSIAESP	FRCLDEFDVF	MDMVNRRIAM	DLILKMADSQ	RFRQFILLTP
1051	QSMSSLPSSK	LIRILRMSDP	ERGQTTLPFR	PVTQEEEDDQ	R

50

BLASTP hits

55

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_11n4, frame 1

SWISSPROT:RA18_SCHPO DNA REPAIR PROTEIN RAD18., N = 1, Score =
1021, P
= 5.2e-103

5

PIR:S51470 hypothetical protein YLR383w - yeast (Saccharomyces
cerevisiae), N = 1, Score = 823, P = 5e-82

10 >SWISSPROT:RA18_SCHPO DNA REPAIR PROTEIN RAD18.
Length = 1,140

HSPs:

15 Score = 1021 (153.2 bits), Expect = 5.2e-103, P = 5.2e-103
Identities = 315/1091 (28%), Positives = 540/1091 (49%)

Query: 2 AKRKEENFSSPKNAKRPRQEELEDF--DKDGDDEDECKGTTTLTAAE----
VGIIIESIHLKN 55

20 A R ++N ++ + +E ++DG+ D T T +

VG+IE IHL N

Sbjct: 45

ASRNQDNRPERQSRLLQRSSSLIEQVRGNEGDGENDVLNQTRETNSTNFDNRVGVIECIHLVN 104

25 Query: 56
FMCHSMLGPXXXXXXXXXXXXXXXXXXAVLTALIVGLGGRVATNRGSSSLKGFVKDQGN 115
FMCH L A+LT L + LG +A TNR ++K

VK G+N

Sbjct: 105 FMCHDSL-

30 KINFGPRIN FVIGHNGSGKSAILTGLTICLGAKASNTNRAPNMKSLVKQGKN 163

Query: 116
SADISITLRNRGDDAFKASVYGNISILIQQHISIDGSRSYKLKSATGSVVSTRKEELIAIL 175
A IS+T+ NRG +A++ +YG SI I++ I +GS Y+L+S

35 G+V+ST+++EL I

Sbjct: 164

YARISVTISNRGFEAYQPEIYGKSITIERTIRREGSSEYRLRSFNGTVISTKRDELDNIC 223

Query: 176
40 DHFNIQVDNPVSVLTQEMSKQFLQSKNEGDKYKFFMKATQLEQMKEDYSYIMETKERTKE 235
DH +Q+DNP+++LTQ+ ++QFL + + +KY+ FMK QL+Q++E+YS I

++ TK

Sbjct: 224

DHMGLQIDNPMNILTQDTARQFLGNSSPKEKYQLFMKGIQLKQLEENYSLIEQSLINTKN 283

45 Query: 236
QIHQGEERLTTELKRQCVKEERFQSIAGLSTMKTNLESCLKHEMAWAVVNEIEKQLNAIRD 295
+ + ++ L ++ E + ++ + LE K EM WA V

E+EK+L

50 Sbjct: 284

VLGNKKTGVSYLAKKEEEYKLLWEQSRATENLHNLLEQKKGEMVWAQVVEVEKEL----- 338

Query: 296 NIKIGEDRAARLDRKMEEQQVRLNEAEQKYKDIDKLEKISEETNARAP-
ECMALKADV 354

55 + E + K+ E + L DI K+ EE RA E

K+

Sbjct: 339 --LLAEKEFQHAEVKLSEAKENLESIVTNQSDIDGKISS-
KEEVIGRAKGETDTTKSKFE 395

Query: 355
 AKKRAYNEAEVLYNRSLSNEYKALKKDDQCKRIEELKKSTDQSLERPERLERQKKISWLK 414
 + ++ Y +N+ K+D + I K D E ER
 5 ++ +
 Sbjct: 396 DIVKTFDG----YRSEMNDVDIQRDIQN---
 SINAAKSCLDVYREQLNTERARENNLGG 448

Query: 415 ERVKAFQNZQNSVNDQIEQF-QQAIEKDKE-----EHG-----
 10 KIKREELDVKHALS 460
 +++ N+ N++ +EI +Q +E + + E G + ++
 + + +S
 Sbjct: 449
 SQIEKRANESNNLQREIADLSEQIVELESKRNDLHSALLEMGGNLTSLLTKKDSIANKIS 508

Query: 461
 15 YNQRQLKELKDSKTDRLKRFGNVPALLEAIDDAYRQGHFTYKPVGPLGACIHLRDPOLA 520
 LK L+D + D++ FG N+P LL+ I R+ F + P GP+G +
 +++ +
 20 Sbjct: 509 DQSEHLKVLEDVQRDKVSAFGKNMPQLLKLIT---
 RETRFQHPKGPMPGKYM TVKEQKWH 565

Query: 521
 25 LAIESCLKGLLQAYCCHNHADERVLQALMKRFYLPGTSRPPIIVSEFRNEIYDVRHRAAY 580
 L IE L ++ + +H D+ +L+ LM++ T ++V +
 YD ++
 Sbjct: 566 LIIERILGNVINGFIVRSHHDQLILKELMRQSNCHAT----VVVGK-----
 YDPFDYSSG 616

Query: 581 HPD--
 30 FPTVLTALIDNAVAVANSLIDMRGIETVLLIKNNSVARAVMQSQKPPKNCREAFT 638
 PD +PTVL ++ D+ V ++LI+ GIE +LLI++ A A M+ +
 N + +
 Sbjct: 617 EPDSQYPTVLKIIKFDDDEVLHTLINHLGIEKMLLIEDRREAEAYMK--
 35 RGIANVTQCYA 674

Query: 639 ADG-DQVFAGRYSSSENTR--PKFLSRDQVDSEI---
 SDLENEVENKTAQILNLQQLHSAL 692
 D ++ + R S++ + K + I S E E K L
 40 Q + ++
 Sbjct: 675
 LDPRNRGYGFRIVSTQRSSGISKVTPWNRPPRIGFSSSTSIEAEKKILDDLKKQYNFASN 734

Query: 693 E-KDIKHNEELLKRCQLHYKELKMKIRKNIS-EIRELENIEEHQ-SV-D---
 45 IATLEDEA 745
 + + K + KR + E I+K I + RE+ ++E + SV D
 I TLE
 Sbjct: 735
 QLNKAKIEQAKFKRDEQLLVEKIEGICKRILLKRREVNLSLESQELSVLDTEKIQTLEIRRI 794

Query: 746 QENKSKMKMVEEHMEQKQENMEH-
 50 LKSLKIEAENKYDAIKFKINQLSELADPLKDELN-L 803
 E + +++ ++ K N EH ++ + + + KI ++
 L+ EL+ L
 55 Sbjct: 795 SETEKELESYAGQLQDAK-
 NEEHRIRDNRQPVIEEIRIYREKIQTETQRLSSLQTELSRL 853

Query: 804

ADSEVDNQKRGKRHYEEKQKEHLDTLNXXXXXXXXXXXXXXXXXSQARQICPERIEVEKS 863
D + +++ +RH + + + L ++A C

ER+ V+ S

5 Sbjct: 854 RDEKRNSEVDIERH-RQTVESCTNILREKEAKKVQCAQVVADYTAKANTRC-
ERVPVQLS 911

Query: 864 ASILDKEINRLRQKIQAHASHG-

DREEIMRQYQEAARETYLDLDSKVRTLKKFIKLLGEI 922

10 + LD EI RL+ +I G E+ Y A+E + V L +
++ L E

Sbjct: 912

PAELDNEIERLQMQIAEWRNRTGVSVQAAEDYLNAAKEKHDQAKVLVARLTQLLQALEET 971

15 Query: 923

MEHRFKTYQQRCLTLRCKLYFDNLLSQRAYCGKMNFDHKNETLSISVQPGEGNKA-AF 981
+ R + + +FR+ +TLR K F+ LSQR + GK+ H+ E L V P

N A A

Sbjct: 972

20 LRRRNEMWTKFRKLITLRTKELFELYLSQRNFTGKLVIKHQEEFLEPRVYPANRNLATAH 1031

Query: 982 N-----

DMRALSGGERSFSTVCFILSLWSIAESPFRCLEDFDVYMDMVNRRIAMDIL 1034

25 N ++ LSGGE+SF+T+C +LS+W P RCLDEFDV+MD VNR

+++ ++

Sbjct: 1032

NRHEKSKVSVQGLSGGEKSFATICMLLSIWEAMSCPLRCLDEFDVFMDAVNRLVSIKMMV 1091

Query: 1035 KMADSQRFRQFILLTPQSMSSLPSSKLIRILRMSDPERGQTTLP 1078

30 A +QFI +TPQ M + K + + R+SDP + LP

Sbjct: 1092 DSAKDSSDKQFIFITPQDMGQIGLDKDVVVFRSLDPVSSSALP 1135

Pedant information for DKFZphamy2_11n4, frame 1

35 -----

Report for DKFZphamy2_11n4.1

40 [LENGTH] 1091

[MW] 126326.13

[pI] 6.57

[HOMOL] SWISSPROT:RA18_SCHPO DNA REPAIR PROTEIN RAD18. 1e-109

45 [FUNCAT] 03.19 recombination and dna repair [S. cerevisiae, YLR383w] 1e-88

[FUNCAT] 08.07 vesicular transport (golgi network, etc.) [S. cerevisiae, YDL058w] 3e-16

50 [FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae, YDL058w] 3e-16

[FUNCAT] 09.13 biogenesis of chromosome structure [S. cerevisiae, YLR086w] 2e-14

[FUNCAT] 1 genome replication, transcription, recombination and repair [M. jannaschii, MJ1643] 3e-14

55 [FUNCAT] 30.04 organization of cytoskeleton [S. cerevisiae, YIL149c] 1e-12

[FUNCAT] 03.22 cell cycle control and mitosis [S. cerevisiae, YDR356w] 8e-12

- [FUNCAT] 09.10 nuclear biogenesis [S. cerevisiae, YDR356w] 8e-12
 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YFL008w] 3e-11
 5 [FUNCAT] 11.04 dna repair (direct repair, base excision repair and nucleotide excision repair) [S. cerevisiae, YKR095w] 2e-09
 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YOR216c] 5e-09
 [FUNCAT] 03.25 cytokinesis [S. cerevisiae, YHR023w MY01 - myosin-1 isoform] 8e-08
 10 [FUNCAT] 03.04 budding, cell polarity and filament formation [S. cerevisiae, YHR023w MY01 - myosin-1 isoform] 8e-08
 [FUNCAT] 08.22 cytoskeleton-dependent transport [S. cerevisiae, YHR023w MY01 - myosin-1 isoform] 8e-08
 15 [FUNCAT] 06.07 protein modification (glycosylation, acylation, myristylation, palmitylation, farnesylation and processing) [S. cerevisiae, YKL201c] 2e-07
 [FUNCAT] 03.13 meiosis [S. cerevisiae, YDR285w] 4e-07
 [FUNCAT] 30.13 organization of chromosome structure [S. cerevisiae, YDR285w] 4e-07
 20 [FUNCAT] 98 classification not yet clear-cut [S. cerevisiae, YJR134c] 7e-07
 [FUNCAT] 06.10 assembly of protein complexes [S. cerevisiae, YPR141c] 7e-07
 25 [FUNCAT] 30.05 organization of centrosome [S. cerevisiae, YPR141c] 7e-07
 [FUNCAT] 11.01 stress response [S. cerevisiae, YPR141c] 7e-07
 [FUNCAT] 03.07 pheromone response, mating-type determination, sex-specific proteins [S. cerevisiae, YPR141c] 7e-07
 30 [FUNCAT] r general function prediction [H. influenzae, HI0756] 1e-06
 [FUNCAT] 10.05.99 other pheromone response activities [S. cerevisiae, YHR158c] 2e-06
 [FUNCAT] 05.04 translation (initiation, elongation and termination) [S. cerevisiae, YAL035w] 3e-04
 35 [FUNCAT] 30.02 organization of plasma membrane [S. cerevisiae, YER008c] 4e-04
 [FUNCAT] 08.16 extracellular transport [S. cerevisiae, YER008c] 4e-04
 40 [FUNCAT] 09.04 biogenesis of cytoskeleton [S. cerevisiae, YKL179c] 7e-04
 [FUNCAT] 03.22.01 cell cycle check point proteins [S. cerevisiae, YGL086w] 7e-04
 [FUNCAT] 08.01 nuclear transport [S. cerevisiae, YDL207w] 0.001
 45 [FUNCAT] 04.07 rna transport [S. cerevisiae, YDL207w] 0.001
 [BLOCKS] BL00326C Tropomyosins proteins
 [BLOCKS] PR01004B
 [BLOCKS] BL00121A Colipase proteins
 [BLOCKS] PF00580A
 50 [SCOP] d2tmab_1.105.4.1.1 Tropomyosin [rabbit (Oryctolagus cuniculus)] 3e-06
 [EC] 3.6.1.32 Myosin ATPase 9e-20
 [PIRKW] phosphotransferase 9e-16
 [PIRKW] nucleus 2e-10
 55 [PIRKW] blocked amino end 2e-07
 [PIRKW] citrulline 2e-10
 [PIRKW] tandem repeat 9e-20
 [PIRKW] heterodimer 3e-11

	[[PIRKW]]	endocytosis 2e-13
	[[PIRKW]]	heart 9e-20
	[[PIRKW]]	polymorphism 1e-10
	[[PIRKW]]	serine/threonine-specific protein kinase 9e-16
5	[[PIRKW]]	transmembrane protein 8e-15
	[[PIRKW]]	zinc finger 2e-13
	[[PIRKW]]	metal binding 2e-13
	[[PIRKW]]	DNA binding 2e-06
	[[PIRKW]]	muscle contraction 9e-20
10	[[PIRKW]]	acetylated amino end 3e-13
	[[PIRKW]]	actin binding 9e-20
	[[PIRKW]]	mitosis 8e-10
	[[PIRKW]]	microtubule binding 3e-09
	[[PIRKW]]	chromosomal protein 3e-11
15	[[PIRKW]]	ATP 9e-20
	[[PIRKW]]	receptor 2e-06
	[[PIRKW]]	thick filament 9e-20
	[[PIRKW]]	phosphoprotein 2e-14
	[[PIRKW]]	glycoprotein 1e-10
20	[[PIRKW]]	skeletal muscle 1e-18
	[[PIRKW]]	calcium binding 2e-10
	[[PIRKW]]	alternative splicing 3e-12
	[[PIRKW]]	DNA condensation 3e-11
	[[PIRKW]]	P-loop 9e-20
25	[[PIRKW]]	coiled coil 9e-20
	[[PIRKW]]	heptad repeat 1e-10
	[[PIRKW]]	methyated amino acid 9e-20
	[[PIRKW]]	basement membrane 1e-10
	[[PIRKW]]	immunoglobulin receptor 4e-09
30	[[PIRKW]]	peripheral membrane protein 2e-13
	[[PIRKW]]	cardiac muscle 9e-20
	[[PIRKW]]	extracellular matrix 1e-10
	[[PIRKW]]	hydrolase 9e-20
	[[PIRKW]]	microtubule 2e-10
35	[[PIRKW]]	muscle 2e-14
	[[PIRKW]]	membrane protein 1e-10
	[[PIRKW]]	EF hand 2e-10
	[[PIRKW]]	cell division 8e-10
	[[PIRKW]]	cytoskeleton 1e-13
40	[[PIRKW]]	hair 2e-10
	[[PIRKW]]	calmodulin binding 2e-13
	[[PIRKW]]	Golgi apparatus 1e-08
	[[PIRKW]]	smooth muscle 2e-07
	[[SUPFAM]]	conserved hypothetical P115 protein 4e-26
45	[[SUPFAM]]	myosin heavy chain 9e-20
	[[SUPFAM]]	unassigned Ser/Thr or Tyr-specific protein kinases 9e-16
	[[SUPFAM]]	centromere protein E 3e-09
	[[SUPFAM]]	calmodulin repeat homology 2e-10
50	[[SUPFAM]]	alpha-actinin actin-binding domain homology 7e-07
	[[SUPFAM]]	myosin motor domain homology 9e-20
	[[SUPFAM]]	tropomyosin 5e-08
	[[SUPFAM]]	plectin 7e-07
	[[SUPFAM]]	pleckstrin repeat homology 3e-09
55	[[SUPFAM]]	trichohyalin 2e-10
	[[SUPFAM]]	hypothetical protein MJ1322 2e-06
	[[SUPFAM]]	ribosomal protein S10 homology 7e-07
	[[SUPFAM]]	protein kinase C zinc-binding repeat homology 3e-09

-64-

```
SEQ  GPNVPALLEAIDDAYRQGHTYKPVGPLGACIHLRDPELALAIESCLKGLLQAYCCHNHA
SEG  .....
PRD  hhhhhhhhhhhhhhhhhhhhhhhcccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
C0ILS
```

[illegible][illegible]

```

SEQ SRDVDSEISDLENEVENKTAQILNLQQHLSALEKD IKHNEELLKR CQLHYKELKM KIRKN
SEG .....
PRD hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
COILS -.-CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC-.....

```

```

30 SEQ ISEIRELENIEEHQSVDIATLEDEAQENKSKMKMVEEHMEQQKENMEHLKSLKIEAENKY
   SEG .....
   PRD hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
   COILS .....CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

```

```

35 SEQ   DAIKFKINQLSELADPLKDELNLADSEVDNQKRGRHYEEKQKEHLDTLNKKKRELDMKE
    SEG   .....xxxxxxxxxxx
    PRD   hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
    COILS CCCCCCCCCCCC.....CCCCCCCCCCCCCCCCCCCC

```

```

40 SEQ KELEEKMSQARQCIPERIEVEKSASILDKKEINRLRQKIQAESHASHGDRREIMRQYQEARE
   SEG xxxxxxxx.....
   PRD hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccccchhhhhhhhhhhhhh
   COILS CCCCCCCCCCCCCCC.....

```

[illegible]

```

55 SEQ      DHKNETLSISVQPGEGNKAAFNDMRALSGGERSFSTVCFILSLWSIAESPFRCLDEFDVY
   SEG      .....
   PRD      eccccccccccccchhhhhhhccccccccchhhhhhhhhhhhhhhhhccccchhhhhhhh
   COILS     .....

```

SEQ MDMVNRRIAMDILKMA DSQRFRQFILLTPQSMSSLPSSKLIRILRMSDPERGQTTL PFR

SEG
 PRD hhhhhhhhhhhhhhhhhhhhhhhceeeeecccccccccccccccccccccccccccccccccc
 COILS

5

SEQ PVTQEEDDDQR
 SEG
 PRD chhhhhhccc
 COILS

10

Prosite for DKFZphamy2_11n4.1

15	PS00016	126->129	RGD	PD0C00016
	PS00017	76->84	ATP_GTP_A	PD0C00017

20

(No Pfam data available for DKFZphamy2_11n4.1)

DKFZphamy2_121f19

5 group: cell structure and motility

DKFZphamy2_121f19 encodes a novel 251 amino acid protein with high similarity to a Rat ankyrin binding glycoprotein-1 related mRNA.

10

Ankyrin binding glycoproteins play a role in neural cell adhesion and in prosate tumor cell transformation. DKFZphamy2_121f19.p3 is expressed in brain, uterus and prostate above average.

15

The new protein can find application modulation of cyto skeleton-membrane interactions.

20

similarity to ankyrin binding glycoprotein-1 related mRNA (Rattus norvegicus)

Sequenced by DKFZ

25

Locus: /map="1"

Insert length: 1498 bp

Poly A stretch at pos. 1479, polyadenylation signal at pos. 1460

30

1 CGGCACCTTC GCCGGCGCCC TCGCCACCCC CAGCCCCGCC CCAGAAGGAG

51 CAGCCCCCGG CGGAGACCCC TACAGACGCT GCTGTCTTGA CCTCACCCCC

101 AGCCCCTGCT CCCCCGGTGA CCCCTAGCAA ACCAATGGCC GGCACCACAG

151 ACCGAGAAGA AGCCACTCGG CTCTTGCTG AGAAGCGGCG CCAGGCCCGG

201 GAGCAGCGGG AGCGCGAGGA GCAGGAGCGG AGGCTGCAGG CAGAAAGGGA

35

251 CAAGCGAATG CGAGAGGAGC AGCTGGCAGG GGAGGCCGAG GCCCGGGCGG

301 AGCGGGAGGC GGAGGCCCGG AGGCGGGAGG AGCAGGAGGC ACAGAGAGAAG

351 GCGCAGGCCG AGCAGGAGGA GCAGGAGCGG CTGCAGAAGC AGAAAGAGGA

401 GGCCGAAGCT CGGTGCGGGG AAGAGGCGGA GCGGCAGCGT CTGGAGCGGG

451 AAAAGCACTT CCAGCAGCAG GAGCAAGAGC GGCAAGAGCG CAGAAAGCGT

40

501 CTGGAGGAGA TCATGAAGAG GACTCGGAAG TCAGAAGTTT CTGAAACCAA

551 GAAGCAGGAC AGCAAGGAGG CCAACGCCAA CGGTTCCAGC CCAGAGCCTG

601 TGAAAGCTGT GGAGGCTCGG TCCCCAGGGC TGCAGAAGGA GGCTGTGCAG

651 AAAGAGGAGC CCATCCCACA GGAGCCTCAG TGGAGTCTCC CAAGCAAGGA

701 GTTGCCAGCG TCCCTGGTGA ATGGCCTGCA GCCTCTCCA GCACACCAGG

45

751 AGAATGGCTT CTCCACCAAC GGACCCTCTG GGGACAAGAG TCTGAGCCGA

801 ACACCAGAGA CACTCCTGCC CTTTGCAGAG GCAGAAGCCT TCCTCAAGAA

851 AGCTGTGGTG CAGTCCCCGC AGGTCACAGA AGTCCTTTAA GAGGGTTTGC

901 CTTGGATCCG GGCACAGTTG TGAGGGCTCC TCTGCATCAC CTACCAGGAT

951 GTCTGGAGGA GAAAAAGACA GAACAAAGAT GGAAGTGGCC TGGGCCCTG

50

1001 GGGGTGGGTC CTCTCTGTTG TTTTAAATCT GCACCTTATA GACTGATGTC

1051 TCTTTGGCCG GAGCCAGATC TGCCCTCAG TGCATTCTG TGCTCGCACG

1101 CGCAGACATC CCTTCTCCCC CATAACACA TATACTCA CAGCCTCTCT

1151 GGCCTCTTCC CTTGGGGAGG GGCCACCTGT AGTATTTGCC TTGATTTGGT

1201 GGGGTACAGT GGATGTGAAT ACTGTAAATA GCTTGTGCTC AGACTCCTCT

55

1251 GCGTGGAGAG GGTGGGTGCA GGAGGCAGAC CCTCCCCCA AAGCCCCCTG

1301 GGGAGATCTT CCTCTCTCTA TTAACTGTA ACTGAGGGGG ATCCCAGGTC

1351 TGGGGATGGG GGACACCTTG GGCCACAGGA TACTGGTTGC TTCAGGGGTA

1401 CCCATGCCCC CTGCCCTCGC CTGGAATCAG TGTTACTGCA TCTGATTAAA

1451 TGTCTCCAGA AATAAAGAAT AATTCTGCCA AAAAAAAAAA AAAAAAAAAA

BLAST Results

5

No BLAST result

10

Medline entries

No Medline entry

15

Peptide information for frame 3

20 ORF from 135 bp to 887 bp; peptide length: 251
 Category: putative protein
 Classification: Cell signaling/communication

25 1 MAGTTDREEA TRLLAEKRRQ AREQREREERQ ERRLQAERDK RMREEQLARE
 51 AEARAEREAE ARRREEQEAR EKAQAEQEEQ ERLQKQKEEA EARSREEAER
 101 QRLEREKHFQ QQEQRERQERR KRLEEIMKRT RKSEVSETKK QDSKEANANG
 151 SSPEPVKAVE ARSPGLQKEA VQKEEPIQAE PQWSLPSKEL PASLVNGLQP
 201 LPAHQENGFS TNGPSGDKSL SRTPETLLPF AEAEAFLLKA VVQSPQVTEV
 251 L

30

BLASTP hits

35 No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_121f19, frame 3

No Alert BLASTP hits found

40

Pedant information for DKFZphamy2_121f19, frame 3

Report for DKFZphamy2_121f19.3

45

50 [LENGTH] 295
 [MW] 33517.96
 [pI] 5.61
 [HOMOL] TREMBLNEW:AB033013_1 gene: "KIAA1187"; product:
 "KIAA1187 protein"; Homo sapiens mRNA for KIAA1187 protein,
 partial cds. 1e-64
 [BLOCKS] PF01140
 55 [BLOCKS] BL00412 Neuromodulin (GAP-43) proteins
 [BLOCKS] BL00826C
 [BLOCKS] BL00422C Granins proteins
 [BLOCKS] PR00167C
 [BLOCKS] PF00992A



5	[KWD]	LOW_COMPLEXITY	51.19 %
	[KWD]	COILED COIL	10.51 %

```

      SEQ GLQLPLAHQENGFTNGPSGDKSLSRTPETLLPFAEAEAFKKAVVQSPQVTEVL
      SEG .....
35 PRD eccccccccccccccccccccccchhhhhhhhhhhhcccccccc
COILS .....

```

40 (No Prosite data available for DKFZphamy2_121f19.3)
(No Pfam data available for DKFZphamy2_121f19.3)

DKFZphamy2_121m2

5 group: cell cycle

DKFZphamy2_121m2 encodes a novel 480 amino acid protein with similarity to human PA2b-T2 protein.

- 10 PA2b-T2 is a p53 responsive gene. The protein is predominantly expressed in brain, breast and kidney and may represent a potential novel regulator of cellular growth. Isoforms are differentially induced by genotoxic stress (UV, gamma-irradiation and cytotoxic drugs) in a p53-dependent manner.

- 15 The new protein can find application in modulating cell division and apoptosis pathways.

- 20 similarity to PA2b nuclear protein isoforms (Homo sapiens)
probably differential polyadenylation

Sequenced by DKFZ

- 25 Locus: unknown

Insert length: 3327 bp

- 30 Poly A stretch at pos. 330b, polyadenylation signal at pos. 3279

1 TCCAGCACCA AAGCGGCCGT TCTCGGATTC CGGAGCGTTC TGGAGCCCCG
51 AGAGACGCCC CGGGGTTCTA GAAGCTCCCC GGCGGCGCCC AGTCCCGGCT
101 TCATTCGGGC GTCCCTCCGA AACCCACTCG GGTGCACGGG TCGTCGGCGA
35 151 GCCGCGACCG GGTCTTGGCG CGCACCATGA TCGTGGCGGA CTCCGAGTGC
201 CGCGCAGAGC TCAAGGACTA CCTGCGGTTT GCCCCGGGCG GCGTCGGCGA
251 CTCGGGCCCC GGAGAGGAGC AGAGGGAGAG CCGGGCTCGG CGAGGCCCTC
301 GAGGGCCAG CGCCTTCATC CCCGTGGAGG AGGTCTTTCG GGAGGGGGCT
351 GAGAGCCTCG AGCAGCACCT GGGGCTGGAG GCACTGATGT CCTCTGGGCG
40 401 AGTAGACAAC CTGGCAGTGG TGATGGGCCT GCACCCTGAC TACTTTACCA
451 GCTTCTGGCG CCTGCACTAC CTGCTGCTGC ACACGGATGG TCCCTTGGCC
501 AGCTCCTGGC GCCACTACAT TGCCATCATG GCTGCCGCCC GCCATCAGTG
551 TTCTTACCTG GTAGGCTCCC ACATGGCCGA GTTTCTGCAG ACTGGTGGTG
601 ACCCTGAGTG GCTGCTGGGC CTCCACCGGG CCCCCGAGAA GTGCGCAAA
45 651 CTCAGCGAGA TCAACAAGTT GCTGGCGCAT CGGCCATGGC TCATCACCAA
701 GGAACACATC CAGGCCTTGC TGAAGACCGG CGAGCACACT TGGTCCCTGG
751 CCGAGCTCAT TCAGGCTCTG GTCCTGCTCA CCCACTGCCA CTCGCTCTCC
801 TCCTTCGTGT TTGGCTGTGG CATCCTCCCT GAGGGGGATG CAGATGGCAG
851 CCCTGCCCCC CAGGCACCTA CACCCCTAG TGAACAGAGC AGCCCCCAA
50 901 GCAGGGACCC GTTGAACAAC TCTGGGGGCT TTGAGTCTGC CCGCGACGTG
951 GAGGCGCTGA TGGAGCGCAT GCAGCAGCTG CAGGAGAGCC TGCTGCGGGA
1001 TGAGGGGACG TCCCAGGAGG AGATGGAGAG CCGCTTTGAG CTGGAGAAAGT
1051 CAGAGAGCCT GCTGGTGACC CCCTCAGCTG ACATCCTGGA GCCCTCTCCA
1101 CACCCAGACA TGCTGTGCTT TGTGGGAAGAC CCTACTTTTC GATATGAGGA
55 1151 CTTCACTCGG AGAGGGGCTT AGGCACCCCC TACCTTCCGG GCCCAGGATT
1201 ATACCTGGGA AGACCATGGC TACTCGCTGA TCCAGCGGCT TTACCCTGAG
1251 GGTGGGCGAG TGCTGGATGA GAAGTTCCAG GCAGCCTATA GCCTCACCTA
1301 CAATACCATC GCCATGCACA GTGGTGTGGA CACCTCCGTG CTCCGAGGG


```

1351 CCATCTGGAA CTATATCCAC TGCCTCTTTG GCATCAGATA TGATGACTAT
1401 GATTATGGGG AGGTGAACCA GCTCCTGGAG CGGAACCTCA AGGTCTATAT
1451 CAAGACAGTG GCCTGCTACC CAGAGAAGAC CACCCGAAGA ATGTACAACC
1501 TCTTCTGGAG GCACTTCCGC CACTCAGAGA AGGTCCACGT GAACTTGCTG
5 1551 CTCCTGGAGG CGCGCATGCA AGCCGCTCTG CTGTACGCCC TCCGTGCCAT
1601 CACCCGCTAC ATGACCTGAC TCCTGAGCAG GACCTGGGCC CGGTTCAGCT
1651 CCCCACAAGG ACTTCTCTGT CTGGAGACAG CCCCAGACCC TTTTGTGTCC
1701 CATGCCCACC CTCCCCACGC TGCAGTGGGC TTGTGTGTGA TGTGCAGTCC
1751 CGAAGCCACA CCCTCCCTTT TCCTCACTGG AATGGACAGT TCATTGCACT
10 1801 GACTCTGGGA TCTCAGCCCT GCTCCTGGGA GCTGGAAGAG CACTTGGAGA
1851 TCCTAAGGGA CCACACCCTT CCTCCTTCCC CTGCCCACAG AGGCAGAGGG
1901 CACAGGAAAG AAGCCGGGCC AAGCTCGGAA TTAATGTGCC ACAAGTGTG
1951 TGGCCTTCCT GAACTGGGAA GTCCCTGGCT GGCCCCGGGG GGAGAGGGGG
2001 AAATGCCTCC GGGACTGACA CTCCAGGCAG CTTTGCCTTC TCTCCCTGT
15 2051 CATTTCAGAG TTTTATTACC TCCTACTTGC CATTACCCA TCAATGTGAA
2101 AGTCAGGGTC ACAGCTGGTC TGTGTGTCCA GTTCCCTAAA AGCCTGTTCT
2151 GTTGGGCAGC CTGAGGCTGT TGCCCGAATC CTAGTTCAGT TTTTGTACTT
2201 CCTTTGCCCT TTTTCCCTTT TCTCCATGCT TAATGGTGTG AGGCGTCAGG
2251 AGAGAGGCCA AGTACATAAA AAAAAAAAAA AGCAGATTAT CTCTAGAGAG
20 2301 TTTGAGCCTT TGCTGGTCAC ATTGCCTTCT GAAGAGGAGG GAGTATTAGA
2351 TTATAAATCC TCTTTATTTT GGTCTTTTAT GCTTGAGGTT CCAACCTGGA
2401 GCCACAGTGT GTGAGAGGAG GAGGAGAGGG AGAATTCTGT TCTCCAGAG
2451 CTGCACCTGC CTCGCAGAGG CCAGCACCCC ACTCTCCTGC CTCCAGTGGC
2501 CCTGCCGCAG ATGTCTCCCA AAAAGTTGAG CTTTTCTAGA TGGCTTAGGT
25 2551 GGCACCATGG CTCAGCAGGA GGGGCGGGAG GCACCAGGGT TCTTGTTTGG
2601 ACCCTGCCCC TGGGCCATGG CCAGGTGACC ATGGCTACAT TGCCAAACCT
2651 CTGACTGCCA CAGCTGCAGA CTGAGAGGGT GGGTCTGAGT CCCCACAATG
2701 TCTGAAGCTG CCCCTGGGAT TCTCAGGCCA ACCTGCCAAC AGCAAGCGGA
2751 TTTTCTTGCA AGATCAGGGA CCCCATTCTT GCAGCCAGTG TCTCCTGGGT
30 2801 GCCTTCTGAG GACTCCCACC CCCATCCCAG TATCTCATCT GTCCCCTCTC
2851 CTGGGGCTTA AGTGGGTTGC TTCCAGGCAG AAGCAGCCAA GGACCGATTTC
2901 CAGGCACTTT CTGTAGCAAA TGA CTGTGAA TTACGACTTC TCTTGCCCTT
2951 CTTCTAGCAG TCTGTGCCTC CTCTCTGACC AGTTTGGAGG GCACTGAAGA
3001 AAGGCAAGGG CCGTGCTGCT GCTGGGCGGG GCAGGAGAGG AGCCTGGCCA
35 3051 GTGTGCCACA TTAAATACCC GTGCAGGCGC GGAGAAGCAA CCGGCACCCC
3101 CTTCCGGCCT GAAAGCCCTC CCTGCAAGAA GGTGTGCAGG AGAGAAGAGG
3151 CCCCGGCATG GGGATCTGGG TTCTAGAGGG CATGTGATGA CTGTAAATGT
3201 TCACTGGGTG GGTAGGGAGT GGTATCCAGT GTTCAAGTGC AGAAATCTTT
3251 GGCTTTGCTA CCAGTTCCAT ATGATGAGAA ATAAACGTTC GCTGAGGTTT
40 3301 TGTTTCATAA AAAAAAAAAA AAAAAAA

```

BLAST Results

45

No BLAST result

Medline entries

50

95024170:

Buckbinder L., Talbott R., Seizinger B.R., Kley N.; Gene regulation by

55 temperature-sensitive p53 mutants: identification of p53 response genes. Proc. Natl. Acad. Sci. U.S.A. 91(22):10640-10644(1994).

9124117:

Velasco-Miguel S, Buckbinder L, Jean P, Gelbert L, Talbott R, Laidlaw J, Seizinger B, Kley N; PA2b, a novel target of the p53 tumor suppressor and member of the GADD family of DNA damage and growth arrest inducible genes. *Oncogene* 1999 Jan 7;18(1):127-37

10

Peptide information for frame 3

15 ORF from 177 bp to 1616 bp; peptide length: 480
Category: strong similarity to known protein
Classification: Cell division

20 1 MIVADSECRA ELKDYLRFP GGVGDSGPGE EQRESRARRG PRGPSAFIPV
51 EEVLREGAES LEQHLGLEAL MSSGRVDNLA VVMGLHPDYF TSFWRHLHYLL
101 LHTDGPLASS WRHYIAIMAA ARHQC SYLVG SHMAEFLQTG GDPEWLLGLH
151 RAPEKLRKLS EINKLLAHRP WLITKEHIQA LLKTGEHTWS LAELIQALVL
201 LTHCHSLSSF VFGCGILPEG DADGSPAPQA PTPPSEQSSP PSRDPLNNSG
251 GFESARDVEA LMERMQQLQE SLLRDEGTSQ EEMESRFELE KSESLVTPS
25 301 ADILEPSPHP DMLCFVEDPT FGYEDFTRRG AQAPPTFRAQ DYTWEDHGYS
351 LIQRLYPEGG QLLDEKFQAA YSLTYNTIAM HSGVDTSVLR RAIWNYIHCV
401 FGIRYDDYDY GEVNQLLERN LKVYIKTVAC YPEKTTRMY NLFWRHFRHS
451 EKVHVNLALL EARMQAAALLY ALRAITRYMT

30

BLASTP hits

No BLASTP hits available

35

Alert BLASTP hits for DKFZphamy2_121m2, frame 3

40 TREMBL:AF033120_1 gene: "PA2b"; product: "p53 regulated PA2b-T2 nuclear protein"; Homo sapiens p53 regulated PA2b-T2 nuclear protein (PA2b) mRNA, complete cds., N = 1, Score = 1377, P = 9.7e-141

45 TREMBL:AF033122_1 gene: "PA2b"; product: "non-p53 regulated PA2b-T1 nuclear protein"; Homo sapiens non-p53 regulated PA2b-T1 nuclear protein (PA2b) mRNA, complete cds., N = 1, Score = 1363, P = 3e-139

50 TREMBL:AF033121_1 gene: "PA2b"; product: "p53 regulated PA2b-T3 nuclear protein"; Homo sapiens p53 regulated PA2b-T3 nuclear protein (PA2b) mRNA, complete cds., N = 1, Score = 1307, P = 2.5e-133

55 >TREMBL:AF033120_1 gene: "PA2b"; product: "p53 regulated PA2b-T2 nuclear

protein"; Homo sapiens p53 regulated PA2b-T2 nuclear
 protein (PA2b) mRNA,
 complete cds.
 Length = 492

HSPs:

Score = 1377 (206.6 bits), Expect = 9.7e-141, P = 9.7e-141
 Identities = 277/471 (58%), Positives = 334/471 (70%)

Query: 22 GVGDSGPGGEEQRESRARRGPR----GPSAFIPVEEVLRGAESLEQH-
 LGLEALMSSGRV 76
 G G G +Q E R PR GPS FIP +E+L+ G+E + H L

++ + GR+

Sbjct: 22
 GCKQCGGGRDQDEELGIRIPRPLGQGPSRFIPEKEILQVGSEDAQMHALFADSFAALGRL 81

Query: 77
 DNLA VMGLHPDYFTSFWRLLHYLLLHTDGPLASSWRHYIAIMAAARHQCSYLVGSHMAEF 136
 DN+ +VM HP Y SF + + LL DGPL +RHYI
 IMAAARHQCSYLV H+ +F

Sbjct: 82
 DNITLVMVFHPQYLESFLKTQHYLLQMDGPLPLHYRHYIGIMAAARHQCSYLVNLHVND 141

Query: 137
 LQGGDPPEWLLGLHRAPEKLRKLSEINKLLAHRPWLTKEHIQALLKTGEHTWSLAELIQ 196
 L GGD P+W L GL AP+KL+ L E+NK+LAHRPWLTKEHI+ LLK
 EH+WSLAEL+

Sbjct: 142
 LHVGGDPKWLNGLENAPQKLQNLGELNKVLAHRPWLTKEHIEGLLKAEHSWSLAELVH 201

Query: 197 ALVLLTHCHSLSSFVFGCGILPEGDADGXXXXXXXXXXXXX-----
 XXXXXXXXRDPLNNS 249
 A+VLLTH HSL+SF FGCGI PE DG

P+N++
 Sbjct: 202
 AVVLLTHYHSLASFTFGCGISPEIHCDGGHTFRPPSVSNYCICDITNGNHSVDMPVNSA 261

Query: 250 GGF--ESARDVEALMERMQQLQESLLRDEG-
 TSQEEMESRFELEKSESLLVTPSADILE 305
 +S +VEALME+M+QLQE RDE SQEEM SRFE+EK ES+ V
 S+D E

Sbjct: 262 ENVSVSDSFFEVEALMEKMRQLQEC--
 RDEEEASQEEMASRFEIEKRESMFVF-SSDDEE 318

Query: 306
 PSPHPDMLCFVEDPTFGYEDFTRRGAQAPPTFRAQDYTWEDHGYSLIQRLYPEGGQLLDE 365
 +P + ED ++GY+DF+R G P TFR QDY WEDHGYSL+
 RLYP+ GQL+DE

Sbjct: 319 VTPARAVSRHFEDTSYGYKDFSRHGMHVP-
 TFRVQDYCWEDHGYSLVNRLYPDVGLIDE 377

Query: 366
 KFQAAYSLTYNTIAMHSGVDTSVLRRRAIWNYIHCVFIRYDDYDYGEVNQLLERNLKVYI 425
 KF AY+LTYNT+AMH
 VDTSLRRRAIWNYIHC+FGIRYDDYDYGE+NQLL+R+ KVYI
 Sbjct: 378
 KFHIAYNLTYNTMAMHKDVTSMRLRRRAIWNYIHC+FGIRYDDYDYGEINQLLDRSFKVYI 437

KTVACYPEKTRRMYNLFWRHFRHSEKVVHVNLLLLLEARMQAALLYALRAITRYMT 480
KTV C PEK T+RMY+ FWR F+HSEKVVHVNLLLL+EARMQA

KT VVCTPEKVTKRMYDSFWRQFKHSEKVHVNLLIEARMQAELLYALRAITRYMT 492

Pedant information for DKFZphamy2_121m2, frame 3

Report for DKFZphamy2_121m2.3

PRD hheeeeeeeeccccchhhhhhhhhhhhhccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccc

(No Prosite data available for DKFZphamy2_121m2.3)

(No Pfam data available for DKFZphamy2_121m2.3)

5

DKFZphamy2_121017

5 group: transmembrane protein

DKFZphamy2_121017 encodes a novel 212 amino acid protein without similarity to known proteins.

10 The novel protein contains 1 transmembrane region.
No informative BLAST results; No predictive prosite, pfam or SCOP motive.

15 The new protein can find application in studying the expression profile of amygdala-specific genes and as a new marker for amygdala cells.

unknown protein

20

Pedant: TRANSMEMBRANE 1

Sequenced by DKFZ

25 Locus: /map="186.6 cR from top of Chr22 linkage group"

Insert length: 2690 bp

Poly A stretch at pos. 2661, polyadenylation signal at pos. 2634

30

```

      1 TGCTGGGAAA AGTGACTGCG ATTCTGAAGA ACCGCTGCCT TGCAAGGTCA
    51 AGGACATTCA GTGGTTGCTG GGGTCCGCAG ACTACTGCCA CCCACTCACC
   101 ATCAACTCTG TTAGCCCAAT TGCCCTGCTG AACAACTGCC TGAATACAGG
   151 CTTTAGGTTC CCCTGGACTC CAGCCAAGGC TGTTCAAGGTG GGACCATGGT
   35 201 GCTCTTTAAG CGTGATCGGA GGAAGACAC ACAGCAGGGC CACCATTTCCA
   251 TGAATGGGAG GTGTACAGAT CACTTTCTCT TTGTGCTCAG TTCTCTTCTG
   301 TCTCCAGCAG CTATATTGGT AAGACTAGTA CCTGCCAGGG AGAGGTGCCC
   351 CCAAGTGAAG GGGTACAGTG GCACCTGGGA AAAGGCACCT GGAAGGTTTC
   401 CATGTGGCCC AGCCCAGCAT GGAAGCAGGG TGGGAACCTC GCTGTGTGCG
   40 451 CAGCCCTCAC TCTACTCAAG TGGCTTTTTG AGAGCCCTGC CATGTCTGTG
   501 TCAGGCCTGT GCTGCTTCAC ACCCTACAGC TGCCTGGGAA AGGCCGGCCA
   551 CGCTCCCTGT CCACACACTC CCTGTCCACA CACTCCCTGT CCACAACCTG
   601 AGCCGGGCCC TCTGCCTATG GGCACCCAAT CCAAGCAGCT GCTCCACCTT
   651 TGTTTGGCAT GGTGATTTGT GTTTTTTCTC TTGGTGCTTA TGTGTGTGGG
   45 701 CTTGGGACGA GTGCTGGTAT GCACTTAGGA CCTTCTTGAT AGCTCCCTGC
   751 ACTTTGGAAC ACGGAGCAGA TGAGAGAGGG TCAGGGGCTT GCCCTCCACC
   801 TTGGACTTGG AAGAAGCCCA CATTGGAGAG GTGAGGACCC CATGGTGGCT
   851 CTAGTGGAAAG ATACGTTAGT CTCCAGCTAA GGAGGATGAG GCGCAGCCCC
   901 AGAGGGAGAC CTCAGTGATA GGGGATCAGG CTACGAAAGT GGGGGAAGGG
   50 951 AGATGCTTTG TACATATTTT GGGGTTATAA TTTCTCTAAA TTTTAGGAGA
  1001 ACGGGTATTG ATTGATAAAA GGGACAGGCA GTAGTGTTCA ACAGTGCATG
  1051 TGAAGGAAAG TTCTGTTTTT CATGGTTTTG ACATTCTTTG GACTGTATTG
  1101 TGA CTGCTGTG CTGGTCCACA TGGTACCCTT TTGGTAAGTA GGCTTCAGTG
  1151 CATACCAGGG TATCACTGGA GATGGGAGTT AGTGAAGGGG TGACTCCCTG
  55 1201 GCCTAGTATA GTGTGACCCT GGGACAACTT AATGTCCTAA AGCATTTTGG
  1251 TGA CTTCTAG GGAATAGCAA AGACCTATTT CATTGTCCCC AGGTAAGTAT
  1301 GTGATGAGCA ATGAGGAGGA GTGGAAAAA AAACCCAGAA AGTGCGGCAG
  1351 GACCAGCCTG ACGCACACGC TCCTGTTGTC ATGGCAGACA GCCGCCTTGG

```

```

1401 GTGGGCACCA CCCTGGCAGT TCCAGCCTGT AGGGGAGTGA AGGGACATGG
1451 CTGAGCTGGG CATGTGCTGA GGTGACTTA GGGAAACAAGC CCTGGGATTG
1501 GACAAAAGGG CCCATGCTGC AGCCACTGAC TGGGGGCAGA GCTCTGGGTG
1551 GAAGAGGGAA GAGATCCTAA TGGAGGCGCC TCCATCTGCA ACCACAGTTG
5 1601 TAAGGCTCAT GGCACCTCTG CTTGGAAAGC ACTGGTTTAG GGACTTAGAG
1651 AGGTAGGCAC AAGGTGGGTC TCCTGGGTAA GGGAAAGCAAG AGCAGACTGT
1701 TGGGCCAACA GGAGAAGCTC CCCAGAGTAG GGGAGAAGGT TGGGGTGTAG
1751 GGCCTTCCAC GTGGAACAGA CAGCCCCCTGT GTCTCTGTCT CTTGGGGACC
1801 TGAGTTTGGG TGGGGTGGCA GTTGGCACAG CGCAGATGCG GTAGAGATGG
10 1851 GAGGAAACCC AGCTCCTCAC TTCCGTGTGC CTCATGCCTT TGCATACACA
1901 AGCACCAAAC CTACTAGGTC TTCTCATTAC CCATGTAAAC CACATGTTAG
1951 ATAAATTTT GCAAGTAGAG GAAAGAAGGA AATAAAACAT CACATTTTGG
2001 TGTCTCTCAG GCTTTCCCCC CCAACTATGG TTTCTTTGCT TTTTGTTTA
2051 ACATAGTTTT GTTGCTGTCT TCTGTAATGA TACAGTTTTG TGCAGCTGTT
15 2101 TTCCTTAGC ATATCGTGGG CATCTCCCTT TATGATTACT AAATATTTTA
2151 TTTTGGAGTG GCTGTGTACT CTCCCATGTA CTAGATGGAC CATTGTGCCA
2201 GTTGCCAATC ACTAATGCTG TTAATACTT TTCAGTTATA AATTGATGAA
2251 TATCTTTGTG CACAGGCTGT TTCCCAATGT CAAGTTATTA GGGTAGACTC
2301 CAGGAGGTGG GATTCTTCAA CTAAAGAATA TGAAAACCTT TGAGGCTTTT
20 2351 ACTACATATT GACAAAATGG TTTCCGGAAT TATTTGTATC CCCTTACACT
2401 GCCACCAGCA AGGATAAACA TGTCCATCTT GCCCGTATTG GGAATTATCA
2451 TCTGGCTAAA TATTTGCTAA TTTGATAATG AAAAAATAGC ATCGTGTTC
2501 AGTTGGCATT TCACTGACTT CTAGCACGGT TGAACATCTT TCATGTGGAG
2551 CGATTGTATT TCCTCCTTTG TGGATTGTCA GTGTCCTTTG CTCTATCTTC
25 2601 TGGGGTCAGA TAAATTTGTA TGAGCTCGGT ATATATTAAG GATATTAACC
2651 TGGTGTGTGT CAAAAAATAA AAAAAAATAA AAAAAAATAA

```

BLAST Results

30

Entry HS1033E15 from database EMBL:
 Human DNA sequence from clone 1033E15 on chromosome 22q13.1-13.2.
 Contains part of a novel gene, ESTs and a GSS.
 35 Score = 5919, P = 5.1e-262, identities = 1187/1195

Entry HSN128A12 from database EMBL:
 Human DNA sequence from cosmid N128A12 on chromosome 22q12-qter.
 contains ESTs, CpG island.
 40 Score = 5038, P = 0.0e+00, identities = 1014/1019

Entry HSB90346 from database EMBL:
 human STS WI-14034.
 45 Score = 1800, P = 1.4e-76, identities = 392/417

Medline entries

50

No Medline entry

55

Peptide information for frame 1

ORF from 196 bp to 831 bp; peptide length: 212

Category: putative protein
Classification: no clue

5 1 MVLFKRDRRE DTQQGHHSMN GRCTDHFLFV LSSLLSPAAIL LVRLVPARER
51 CPQVKGYSGT WEKAPGRFPC GPAQHGSRVG TLLCRQPSLY SSGFLRALPC
101 LCQACAASHPTAAWERPATL PVHTLPVHTL PVHNCSTRALC LWAPNPSSCS
151 TFVWHGDLCF FSWCLCVWAW DECWYALRTF LIAPCTLEHG ADERGSACGP
201 PPWTWKKPTL ER

10

BLASTP hits

No BLASTP hits available

15

Alert BLASTP hits for DKFZphamy2_121017, frame 1

No Alert BLASTP hits found

20

Pedant information for DKFZphamy2_121017, frame 1

Report for DKFZphamy2_121017.1

25

[LENGTH] 212
[MW] 23727.55
[pI] 8.73
[KW] TRANSMEMBRANE 1

30

SEQ MVLFKRDRREDTQQGHHSMNGRCTDHFLFVLSSLLSPAAILVRLVPARERCPQVKGYSGT
PRD cccchhhhhccccccccccccccccccccchhhhhhhccccceeecccccccccccccccccc
MEMMMMMMMMMMMMMMMMM.....

35

SEQ WEKAPGRFPCGPAQHGSRVGTLLCRQPSLYSSGFLRALPCLCQACAASHPTAAWERPATL
PRD cccccccccccccccccccccceeeccccccccccccccccccccchhhhhhhcccccccccccccc
MEMcc

40

SEQ PVHTLPVHTLPVHNCSTRALC LWAPNPSSCSTFVWHGDL CFFSWCLCVWAWDECWYALRTF
PRD cccccccccccccccccccccceeeccccccccccccccccccccceeeccccccccccccccccccchhhhhhhhe
MEMcc

45

SEQ LIAPCTLEHGADERGSACPPPWTWKKPTLER
PRD eeeeecccccccccccccccccccccccccccccccccccccc
MEMcc

50

(No Prosite data available for DKFZphamy2_121017.1)

(No Pfam data available for DKFZphamy2_121017.1)

DKFZphamy2_12d7

5 group: signal transduction

DKFZphamy2_12d7 encodes a novel 552 amino acid protein, which is a so far unknown alternative spliced form of disks large homolog DLG2.

10

It seems to be predominantly expressed in the retina, germ cells and brain. It contains a SH3-domain and a guanylate kinase domain. These conserved regions are shared among members of the discs-large family of proteins that include human p55, a membrane protein expressed in erythrocytes, rat PSD-95/SAP90, a synapse protein expressed in brain, Drosophila dIg-A, a septate junction protein expressed in various epithelia, and human and mouse ZO-1 and canine ZO-2, two tight junction proteins. The Homologue of Drosophila, dIg-A, acts as a tumor suppressor. All members of this family may be involved in signal transduction.

20

The new protein can find application in modulating/blocking intracellular signal transduction pathways.

25

similarity to disks large homolog DLG2 (Homo sapiens)

alternative splicing: see DLG2
complete cds.

30

frame shift: around position 1437 one C too many

Sequenced by EMBL

Locus: /map="338.6 cR from top of Chr17 linkage group"

35

Insert length: 4220 bp

Poly A stretch at pos. 4180, polyadenylation signal at pos. 4165

```

40      1 CCCGGCTGCG CTGGAGCCGC CCGGAGCTAG GGGCTTCCCG GGGCGCAGGA
      51 GAGACGTTTC AGAGCCCTTG CCTCCTTCAC CATGCCGGTT GCCGCCACCA
     101 ACTCTGAAAC TGCCATGCAG CAAGTCCTGG ACAACTTGGG ATCCCTCCCC
     151 AGTGCCACGG GGGCTGCAGA GCTGGACCTG ATCTTCCTTC GAGGCATTAT
     201 GGAAAGTCCC ATAGTAAGAT CCCTGGCCAA GGCCCATGAG AGGCTGGAGG
     45  251 AGACGAAGCT GGAGGCCGTG AGAGACAACA ACCTGGAGCT GGTGCAGGAG
     301 ATCCTGCGGG ACCTGGCGCA GCTGGCTGAG CAGAGCAGCA CAGCCGCCGA
     351 GCTGGCCAC ACCTCCAGG AGCCCCACTT CCAGTCCCTC CTGGAGACGC
     401 ACGACTCTGT GGCCTCAAAG ACCTATGAGA CACCACCCCC CAGCCCTGGC
     451 CTGGACCCTA CGTTCAGCAA CCAGCCTGTA CCTCCCGATG CTGTGCGCAT
     50  501 GGTGGGCATC CGCAAGACAG CCGGAGAACA TCTGGGTGTA ACGTTCCGCG
     551 TGGAGGGCGG CGAGCTGGTG ATCGCGCGCA TTCTGCATGG GGGCATGGTG
     601 GCTCAGCAAG GCCTGCTGCA TGTGGGTGAC ATCATCAAGG AGGTGAACGG
     651 GCAGCCAGTG GGCAGTGACC CCCGCGCACT GCAGGAGCTC CTGCGCAATG
     701 CCAGTGGCAG TGTATCCTC AAGATCCTGC CCAGCTACCA GGAGCCCCAT
     55  751 CTGCCCCGCC AGGTATTTGT GAAATGTAC TTTGACTATG ACCCGGCCCC
     801 AGACAGCCTC ATCCCCTGCA AGGAAGCAGG CCTGCGCTTC AACCGCGGGG
     851 ACTTGCTCCA GATCGTAAAC CAGGATGATG CCAACTGGTG GCAGGCATGC
     901 CATGTCGAAG GGGGCGAGTG TGGGCTCATT CCCAGCCAGC TGCTGGAGGA
  
```

951 GAAGCGGAAA GCATTTGTCA AGAGGGACCT GGAGCTGACA CCAAACCTCAG
1001 GGACCCTATG CGGCAGCCTT TCAGGAAAGA AAAAGAAGCG AATGATGTAT
1051 TTGACCACCA AGAATGCAGA GTTTGACCGT CATGAGCTGC TCATTTATGA
1101 GGAGGTGGCC CGCATGCCCC CGTTCCGCCG GAAAACCCTG GTACTGATTG
5 1151 GGGCTCAGGG CGTGGGACGG CGCAGCCTGA AGAACAAGCT CATCATGTGG
1201 GATCCAGATC GCTATGGCAC CACGGTGCCC TACACCTCCC GCGGCGCGAA
1251 AGACTCAGAG CGGGAAGGTC AGGGTTACAG CTTTGTGTCC CGTGGGGAGA
1301 TGGAGGCTGA CGTCCGTGCT GGGCGCTACC TGGAGCATGG CGAATACGAG
1351 GGCAACCTGT ATGGCACACG TATTGACTCC ATCCGGGGCG TGGTCGCTGC
10 1401 TGGGAAGGTG TGCGTGCTGG ATGTCAACCC CCAGGCCGGT GAAGGTGCTA
1451 CGAACGGCCG AGTTTGTCCC TTACGTGGTG TTCATCGAGG CCCCAGACTT
1501 CGAGACCTTG CGGGCCATGA ACAGGGCTGC GCTGGAGAGT GGAATATCCA
1551 CCAAGCAGCT CACGGAGGCG GACCTGAGAC GGACAGTGGG GGAGAGCAGC
1601 CGCATCCAGC GGGGCTACGG GCACTACTTT GACCTCTGCC TGGTCAATAG
15 1651 CAACCTGGAG AGGACCTTCC GCGAGCTCCA GACAGCCATG GAGAAGCTAC
1701 GGACAGAGCC CCAGTGGGTG CCTGTCAAGT GGGTGTACTG AGCCTGTTCA
1751 CCTGGTCCTT GGCTCACTCT GTGTTGAAAC CCAGAACCTG AATCCATCCC
1801 CCTCCTGACC TGTGACCCCC TGCCACAATC CTTAGCCCCC ATATCTGGCT
1851 GTCCTTGGGT AACAGCTCCC AGCAGGCCCT AAGTCTGGCT TCAGCACAGA
20 1901 GGCCTGCACG GCCAGGGAGG TGGGCATTCA TGGGGTACCT TGTGCCCAGG
1951 TGCTGCCAC TCCGTGATGCC CATTGGTCAC CAGATATCTC TGAGGGCCAA
2001 GCTATGCCCA GGAATGTGTC AGAGTCACCT CCATAATGGT CAGTACAGAG
2051 AAGAGAAAAG CTGCTTTGGG ACCACATGGT CAGTAGGCAC ACTGCCCTG
2101 CCACCCCTCC CCAGTCACCA GTTCTCCTCT GGAAGTGGCC CACCCACCCC
25 2151 ATTCTGGAC TCCTCCACCC TCTCACCCCT GTGTGCGAGG AACAGGCCCT
2201 GGGCTGTTTC CGTGTGACCA GGGGAATGTG TGGCCCGCTG GCAGCCAGGC
2251 AGGCCCCGGT GGTGGTGCCA GCCTGGTGCC ATCTTGAAGG CTGGAGGAGT
2301 CAGAGTGAGA GCCAGTGCC ACAGCTGCAG AGCACTGCAG CTCCCAGCTC
2351 CTTTGGAAAG GGACAGGGTC CAGGGGCAGA TGCTGCTCGG TCCTTCCCTC
30 2401 ATCCACAGCT TCTCACTGCC GAAGTTTCTC CAGATTTCTC CAATGTGTCC
2451 TGACAGGTCA GCCCTGCTCC CCACAGGGCC AGGCTGGCAG GGGCCATTGG
2501 GCTCAGCCCA GGTAGGGGCA GGATGGAGGG CTGAGCCCTG TGACAACCTG
2551 CTGTTACCAA CTGAAGAGCC CCAAGCTCTC CATGGCCAC AGCAGGCACA
2601 GGTCTGAGCT CTATGTCTTT GACCTTGGTC CATTTGGTTT TCTGTCTAGC
35 2651 CAGGTCCAGG TAGCCCACTT GCATCAGGGC TGCTGGGTTG GAGGGGCTAA
2701 GGAGGAGTGC AGAGGGGACC TTGGGAGCCT GGGCTTGAAG GACAGTTGCC
2751 CTCCAGGAGG TTCCTCACAC ACAACTCCAG AGGCGCCATT TACACTGTAG
2801 TCTGTACAAC CTGTGGTTCC ACGTGCATGT TCGGCACCTG TCTGTGCCTC
2851 TGGCACCAGG TTGTGTGTGT GTGCGTGTGC ACGTGCCTGT GTGTGTGTGT
40 2901 GTGTCAAGTT TAGTTTGGGG AGGAAGCAAA GGGTTTTGTT TTGGAGGTCA
2951 CTCTTTGGGG CCCCTTTCTG GGGGTTCCTC ATCAGCCCTC ATTTCTTATA
3001 ATACCCTGAT CCCAGACTCC AAAGCCCTGG TCCTTTCCTG ATGTCTCCTC
3051 CCTTGTCTTA TTGTCCCCCT ACCCTAAATG CCCCCCTGCC ATAACCTGGG
3101 GAGGGCAGTT TTGTAAAATA GGAGACTCCC TTTAAGAAAG AATGCTGTCC
45 3151 TAGATGTACT TGGGCATCTC ATCCTTCATT ATTCTCTGCA TTCTTCCGG
3201 GGGGAGCCTG TCCTCAGAGG GGACAACCTG TGACACCCTG AGTCCAAACC
3251 CTTGTGCCTC CCAGTTCTTC CAAGTGTCTA ACTAGTCTTC GCTGCAGCGT
3301 CAGCCAAAGC TGGCCCCCTG ACCACTGTGT GCCCATTTC TAGGGAAGGG
3351 GAAGGAGAAT AAACAGAATA TTTATTACAA ATGTTAGAAT ATATTTCTTA
50 3401 TACTAGGAAT CTCATTTGCA TTTGCATAGA CTATACACAT GGGGTGGAAA
3451 GGCCAGGCCCT GCCCCATCT CGTTGGTGTG GCTCTGCGTA TACTACACAC
3501 TCATTCTCCT GCTCCTCTTT TCCCTTAGTC AGTGTCTTT CATCCTGATT
3551 CAGCTCTGCC TTGCATCACC CTCAGCCTAA GGGAGTGGGA AGGAAATGGG
3601 GTGTTTTCTT GCTGACCTGA GGCTATAGGG TCACTTGCCA TTTCTACCT
55 3651 TCTCTGGGGG ATTTGAGGGT AGAGGCAGGG GAAGATCTGT TGTGCAAGTT
3701 GCTTCTGCCC CCTTGATCCA AATGACCATC ATCTCTGATG GAGATGGGT
3751 GGGTACCTGG CCTTCATGGC ACCTTCACTG CTAGGGATGC TCAAGGGGCA
3801 GGCTGGGGC CCTTCCCTCC TGTCTCTTCT CGGTCTTTCC TCTCTGAGCA

3851 GCCTCCTACC TCCCCTGCCT GAGCCCTCAC TCCACAGCCC TCCCAGGTAC
 3901 CTAGCAGAGG CTGTCAGTCC TTGGCTCACC TGGAACAGGG CTGGGGCTGG
 3951 GTTGGAAACAG GTGTGTGCCC CCACCACAGC TCTATGACTC TGTTCTCCCT
 4001 CCCTGCCATT GTGGACTCTT GTATTTGAGG GACCTCAAGA GAGTGAGGAC
 5 4051 CCTACCATCC ACTGTCCATA TTCAGTCCCA GCCCCAGTGC GCTTCCTCTG
 4101 TTCCCTCCCT CAGCCATCCA ATTCTTGAGT TTTCTCACTG ATTGGTTTTTC
 4151 TTTCTTTTTC CTTGGATTAA ATGTGAAAGC AAAGAAAAAA AAAAAAAAAA
 4201 AAAAAAAAAA AAAAAAAAAA

10

BLAST Results

No BLAST result

15

Medline entries

20 96070428:

Mazoyer S, Gayther SA, Nagai MA, Smith SA, Dunning A, van
 Rensburg EJ,
 Albertsen H, White R,
 Ponder BA. A gene (DLG2) located at 17q12-q21 encodes a new
 25 homologue
 of
 the Drosophila tumor suppressor dIg-A. Genomics 1995 Jul
 1;28(1):25-31

30

Peptide information for frame 1

35

ORF from 82 bp to 1437 bp; peptide length: 452
 Category: strong similarity to known protein
 Classification: Cell signaling/communication
 Prosite motifs: GUANYLATE_KINASE_1 (385-402)

40

1 MPVAATNSET AMQQVLDNLG SLPSATGAEE LDLIFLRGIM ESPIVRSLAK
 51 AHERLEETKL EAVRDNNLEL VQEIIRDLAQ LAEQSSTAAE LAHILQEPHF
 101 QSLLETHDSV ASKTYETPPP SPGLDPTFSN QPVPPDAVRM VGIRKTAGEH
 45 151 LGVTFRVEGG ELVIARILHG GMVAQQGLLH VGDIIKEVNG QPVGSDPRAL
 201 QELLRNASGS VILKILPSYQ EPHLPRQVFV KCHFDDYDPAR DSLIPCKEAG
 251 LRFNAGDLLQ IVNQDDANWW QACHVEGGS AGLIPSQLEE KRKAFVKRDL
 301 ELTPNSGTLC GSLSGKKKKR MMYLTTKNAE FDRHELLIYE EVARMPPFRR
 351 KTLVLIGAQQ VGRRLKKNKL IMWDPDRYGT TVPYTSRRPK DSEREGQGYS
 50 401 FVSRGEMEAD VRAGRYLEHG EYEGNLYGTR IDSIRGVVAA GKVCVLDVNP
 451 QA

55

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_12d7, frame 1

No Alert BLASTP hits found

5

Peptide information for frame 2

ORF from 1439 bp to 1738 bp; peptide length: 100

10 Category: strong similarity to known protein

Classification: Cell signaling/communication

Prosites motifs: LEUCINE_ZIPPER (66-87)

15 1 VKVLRTAEFV PYVVFIEAPD FETLRAMNRA ALESGISTKQ LTEADLRRTV
51 EESSRIQRGY GHYFDLCLVN SNLERTFREL QTAMEKLRT E PQWVPVSWVY

20

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_12d7, frame 2

25

No Alert BLASTP hits found

Pedant information for DKFZphamy2_12d7, frame 1

30

Report for DKFZphamy2_12d7.1

35 [LENGTH] 516
[MW] 56458.36
[PI] 6.21
[HOMOL] PIR:A57653 disks large homolog DLG2 - human 0.0
[FUNCAT] 01.03.99 other nucleotide-metabolism activities [S. cerevisiae, YDR454c] 7e-15
40 [FUNCAT] f nucleotide metabolism and transport [H. influenzae, HI1743] 3e-07
[BLOCKS] PR00834F
[BLOCKS] BL00856C
[BLOCKS] BL00856B Guanylate kinase proteins
45 [BLOCKS] BL00856A Guanylate kinase proteins
[SCOP] dlgy_ 3.29.1.1.1 Guanylate kinase [baker's yeast (Saccharomyce 8e-45
[SCOP] dlkwab_ 2.26.1.1.2 Cask/Lin-2 [Human (Homo sapiens) 4e-34
50 [EC] 2.7.4.8 Guanylate kinase 8e-17
[PIRKW] blocked amino end 8e-17
[PIRKW] phosphotransferase 8e-17
[PIRKW] monomer 8e-17
[PIRKW] duplication 5e-29
55 [PIRKW] signal transduction 3e-24
[PIRKW] alternative splicing 5e-29
[PIRKW] P-loop 8e-17
[PIRKW] acetylated amino end 1e-16

[[PIRKW]] membrane protein 9e-74
 [[PIRKW]] magnesium 8e-17
 [[PIRKW]] ATP 8e-17
 5 [[SUPFAM]] SH3 homology 9e-74
 [[SUPFAM]] discs-large tumor suppressor 3e-24
 [[SUPFAM]] unassigned Ser/Thr or Tyr-specific protein kinases 5e-11
 [[SUPFAM]] protein kinase homology 5e-11
 10 [[SUPFAM]] GLGF domain homology 9e-74
 [[SUPFAM]] guanylate kinase 8e-17
 [[SUPFAM]] guanylate kinase homology 9e-74
 [[PROSITE]] GUANYLATE_KINASE_1 1
 [[PFAM]] Src homology domain 3
 15 [[KW]] Irregular
 [[KW]] 3D

SEQ MPVAATNSETAMQQVLDNLGSLPSATGAAELDLIFLRGIMESPIVRSLAKAHERLEETKL
 20 lgky-
 SEQ EAVRDNNLELVQEILRDLAQLAEQSSTAELAHLQEPHFQSLLETHDSVASKTYETPPP
 lgky-
 25 SEQ SPGLDPTFSNQPVPPDAVRMVGIRKTAGEHLGVTFRVEGGELVIARILHGGMVAQQGLLH
 lgky-
 30 SEQ VGDIIKEVNGQPVGSDPRALQELLRNASGSVILKILPSYQEPHLPRQVFVKCHFDDYDPA
 lgky-
 35 SEQ DSLIPCKEAGLRFNAGDLLQIVNQDDANWWQACHVEGGSAGLIPSQLLLEEKRAKAFVKRDL
 lgky-
 SEQ ELTPNSGTLGSLSGKKKKRMMYLTTKNAEFDRHELLIYEEVARMPPFRKTLVLIGAQG
 40 lgky-CCEEEECTTT
 SEQ VGRRLKKNKLIMWDPDRYGTTPYTSRRPKDSEREGQGYSFVSRGEMEADVRAGRYLEHG
 lgky-
 45 TCHHHHHHHHHHTTTTTEEECCCEEECCCCTTTTTTTTTTTEECCHHHHHHHHHHCCEEEEE
 SEQ EYEGNLYGTRIDSIRGVVAAGKVCVLDVNPQAGEGATNGRVCPLRGVHRGPRLRDPAGHE
 lgky-
 EETTEEEEEHHHHHHHHHHHCCEEEECCHH.....
 50 SEQ QGCAGEWNIHQAAHGGGPETDSGGEQPHPAGLRALL
 lgky-

55 Prosite for DKFZphamy2_12d7.1

PS00856

385->403

GUANYLATE_KINASE_1

PD0C00670

Pfam for DKFZphamy2_12d7.1

5 HMM_NAME Src homology domain 3

HMM

10 *pyVIALYDYqAqd.....pDELSFkEGDIIiIIIEdsDD-WWrgRnnn
 +V+ +DY++ + + L F GD ++I++++D+ WW +

Query 228
 VFVKCHFDDYDPARDSLIPCKEAGLRFNAGDLLQIVNQDDANWWQACHVE 276

HMM TNGQEGWIPSNYVEPi*
 ++ G+IPS +E+

15 Query 277 GG-SAGLIPSQLEEK 291

20 Pedant information for DKFZphamy2_12d7.1 frame 2

Report for DKFZphamy2_12d7.2

25 [LENGTH] 175
 [MW] 19721.90
 [pI] 9.69

30 [HOMOL] PIR:A57653 disks large homolog DLG2 - human 7e-53
 [PIRKW] membrane protein 1e-13
 [SUPFAM] SH3 homology 1e-13
 [SUPFAM] GLGF domain homology 1e-13
 [SUPFAM] guanylate kinase homology 1e-13
 [PROSITE] LEUCINE_ZIPPER 1

35 [KW] Alpha_Beta

SEQ MAPRCPTPPGGRKTQSGKVRVTALCPVGRWRLTSVLGATWSMANTRATCMHVLTPSGAW
 PRD ccc

40 SEQ SLLGRCACWMSTPRPVKVLRTAEFVPYVVFIEAPDFETLRAMNRAALESISTKQLTEAD
 PRD ccc

45 SEQ LRRTVEESSRIQRGYGHYFDLCLVNSNLERTFRELQTAMEKLRTEPQWVPVSWVY
 PRD hhh

Prosite for DKFZphamy2_12d7.2

50 PS00029 141->163 LEUCINE_ZIPPER PD000029

55 (No Pfam data available for DKFZphamy2_12d7.2)

5 group: amygdala derived

DKFZphamy2_12g7 encodes a novel 254 amino acid protein without similarity to known proteins.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motif.

The new protein can find application in studying the expression profile of amygdala-specific genes.

15

putative protein

Sequenced by EMBL

20

Locus: unknown

Insert length: 1257 bp

No poly A stretch found, no polyadenylation signal found

25

```

      1 CTCCAAGACT TCCTTGCTGT GAGGCTCGTG TGGACCCCAG AGCATGCACA
    51 GGCTGTTTAC TCCACAGAGT GGCTTTGAGA ATCAGATGAG ACTGTGCTGG
   101 CGAAGGCCCT GTGGGAATGA GGAACGCTGT AGTGTTTGCT GGTCCCTGTT
   30 151 TCTGCCCCCA GGAAAGCAGC TGTGTGAGGA GGAGCGCCGG GCCATGCAGG
   201 CTGCCCTGGA CTCCGTCGTC TGCCACACGC CCCTCAACAA CTTGGCTTTT
   251 TCCCGGAAGG GCAGCGCGCT CACCTTCAGT GTGGCCTTCC AGGCTCTGAG
   301 GACGGGGGCTC TTCGAGCTAA GCCAGCACAT GAAACTGAAG CTGCAGTTCA
   351 CCGCCAGCGT GTCCACCCCT CCACCCGAGG CCCGGCCCCCT CTCCCGCAAG
   401 AGCAGCCCCA GAAGCCCTGC TGTCCGGGAC TTGGTGGAGA GGCATCAGGC
   451 TAGCCTGGGC CGCTCCCACT CTTTCTCCCA CCAGCAGCCT TCCCGAAGCC
   501 ACCTCATGAG GTCGGGCGAG GTGATGGAGC GCAGAGCATC ACGCCCCCTG
   551 TGGCCTCTCC TGTGCGCCGC CCCCTCTACC TGCCCCCGGA CAAGGCTGTG
   601 TTGTCTCTGG ACAAGATTGC CAAGCGCGAG TGCAAGGTCC TGGTGGTGGA
   40 651 ACCCGTCAAG TAGCACCGTG CCAGCTCTGT TCCCTCTTAC ACTCCAGAGA
   701 CCCAACGCCC CCAGAGGGTA TCCTTGCTCC CGGGCTGTGC CTCCCCTGGG
   751 ATGCCTCCCA GACGGGGGTG AAGAGGCCTG GCAGAGCTGC CTGTCTTGTG
   801 TCTGCTGATG AGGGATGGGG GAAGAAGCTG TGAAGTGGGC GGGCATGGCT
   851 GGGACTAAGC CACCAGTATT CCCCACGTT CCTGTGGGGG GGGCTGGCCC
   45 901 ACCCCTAGGC CAGGGCAAGG GTTCCAGAG CTCCCTTGTC CCCGGCCCTT
   951 TACCCTGGTT CTGAGTTTAC AAAGTCTCTT CCTCATTCCC GTTGAGTTCT
  1001 TTCCACCTC TGACATTCCC TCCCTCCCTC CCGCAGGCTG AGATTAGAGG
  1051 GTGGTGATGG CTAAGGGCCC CTGACAGTGA CCTTCCTGTC TCAGGGGTTG
  1101 GGGACAGGGC CAGGTAGCCT CTGCCCCCTT ATGTTTACGT TTGCAGCCTG
   50 1151 AAGCACTTTA ATTTTTTTTT TTTTGGTCT GTCCCTGTAA CTAATTTTCC
  1201 AACTATTGCT TCCAAC TGAA ATAAGACTAT TAAATGCCTG TTCAGAGGGA
  1251 AAAAAAA
```

55

BLAST Results

No BLAST result

No Medline entry

Peptide information for frame 2

Category: putative protein

Classification: no clue

20

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_12g7, frame 2

No Alert BLASTP hits found

35

Report for DKFZphamy2_12g7.2

40

50

55

SEQ PRTRLCCLWTRLPSASARSWWNPSSSTVPALFPLTLQRPNAPRGYPSCRAVPPLGCLPD
SEG
PRD ccc

5

SEQ GGEEAWQSLSCVC
SEG
PRD cchhhhhhhhhccc

10

(No Prosite data available for DKFZphamy2_12g7.2)

(No Pfam data available for DKFZphamy2_12g7.2)

DKFZphamy2_12i1

5 group: amygdala derived

DKFZphamy2_12i1 encodes a novel 283 amino acid protein with weak similarity to F41E6.3 of *Caenorhabditis elegans*.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motive.

The new protein can find application in studying the expression profile of amygdala-specific genes.

15

putative protein

Sequenced by EMBL

20

Locus: /map="3"

Insert length: 2528 bp

Poly A stretch at pos. 2515, polyadenylation signal at pos. 2491

25

```
1 ATATAGTTGG ATCAAACAAA AACAAACACAA TTTGTCCCGA TAATTATCAA
51 ACAGCACAGC TACTTGCCCTT AATTTTAGAG TTAATCACAT TTTGTGTGGA
101 ACATCACACA TATCACATAA AAAACTATAT TATGAACAAG GACTTGCTAA
30 151 GAAGAGTCTT GGTCTTGATG AATTCAAAGC ACACTTTTCT GGCCTTGTGT
201 GCCCTTCGCT TTATGAGGCG GATAATTGGA CTTAAAGATG AATTTTATAA
251 TCGTTACATC ACCAAGGGAA ATCTTTTGA GCCAGTTATA AATGCACTTC
301 TGGATAATGG AACTCGGTAT AATCTGTTGA ATTCAGCTGT TATTGAGTTG
351 TTTGAATTTA TAAGAGTGGA AGATATCAAG TCTCTTACTG CCCATATAGT
35 401 TGAAGAACTTT TATAAAGCAC TTGAATCGAT TGAATATGTT CAGACATTCA
451 AAGGATTGAA GACTAAATAT GAGCAAGAAA AAGACAGACA AAATCAGAAA
501 CTGAACAGTG TACCATCTAT ATTGCGTAGT AACAGATTTT GCAGAGATGC
551 AAAAGCCTTG GAAGAGGATG AAGAAATGTG GTTTAATGAA GATGAAGAAG
601 AGGAAGGAAA AGCAGTTGTG GCACCAGTGG AAAAACCTAA GCCAGAAGAT
40 651 GATTTTCCAG ATAATTATGA AAAGTTTATG GAGACTAAAA AAGCAAAAAG
701 AAGTGAAGAC AAGGAAAACC TTCCCAAAAG GACATCTCCT GGTGGCTTCA
751 AATTTACTTT CTCCCACTCT GCCAGTGCTG CTAATGGAAC AAACAGTAAA
801 TCTGTAGTGG CTCAGATACC ACCAGCAACT TCTAATGGAT CCTCTTCCAA
851 AACCACAAAC TTGCCTACGT CAGTAACAGC CACCAAGGGA AGTTTGGTTG
45 901 GCTTAGTGGA TTATCCAGAT GATGAAGAGG AAGATGAAGA AGAAGAATCG
951 TCCCCCAGGA AAAGACCTCG TCTTGCTCA TAAATATTT ATTAGGGGAC
1001 CCTCAACATG TGGTCTTACA ATGCTGCAAC TGTTCAAGTGA GCTGAAAATC
1051 TGAATCAGAA AGCTTTCTCA ATTGAACCTA TAAATATAC AAGGAGTAGC
1101 AAAAGACAGT ATATCAGCTA AGAGAGTTTA GTTCTAATAA AAATCAGGCT
50 1151 TCCCAGGAAC TTGATTGCTT GCTAGTAATT AAGGGGTTTG CTTTTAGGC
1201 TGTCAAAACA AACATTAGTA ACCAGAACCT GGGAGATAGC TTCTCAGCAA
1251 GGAAAAGTCA CAGGTTTGGG GACGGTTTAG GGGAGGGGAA AAGGTTGATA
1301 TAATAATGCA GGGTTGCTCC TCGGGGTGTC GATCTAGAAA CAATTTTACA
1351 GAAC TTCAGT TGTA AACTCA ATAACATTAC TTGTATAATG GTGCTGGCCA
55 1401 TGTTGTTGTT TTAATCAGTT GCCTCTTTTT AAAAGAAATT TTTATGGAAA
1451 ACACATTCAA CTATCATTA AAAAAATGAAG TTAAGCTGTT GGGACCATTT
1501 CTTTAAGATT TAACAAAAGT TCAGCCTTTT AGGTAGTTGA AGGGAAGTAC
1551 ACCCCGTATT CAGCACATGT TGAGTTTTCT ACACCAGGAA TTTTCAATAT
```

1601 GTATATTGAT GAAAACAAGC TCAATTCAAA CTGGACAGTT TTAAGATAAT
 1651 GTTAAAATCA GCACTTTTAT AGACAACGAA GGCCAAGAAT CAGTACAGTA
 1701 GTATTTCAAA ATGATTTTCT CTAGAAATTT GAAAGTAGAT CGAACAGAAT
 1751 GTTGTCAACC GCCTACCAAGT ACAATCTTTT GTGGGAAGATA CTTTGAAATC
 5 1801 ACTTTCTACT TTGTTAGTAA AGTTCTGTCT TTCCAGAGCT GCAAGTTTTA
 1851 AAGTGTACT TATACAGACC AACCAAGAAT AGTGCTGAAT TAAGTGGCAT
 1901 TTAGTATCTA GAAGCCATTT TGATCCAAGA AGCTACTTAA GTGTCAAAGT
 1951 CAGCATGCAG CACATGTAGC TTTTCTGTAA ACAAGGGTGT GATATGAAAG
 2001 CTGCTTTTTT AAGAAGAGTA AAAGCACATT CCATATACGT AAGTGAATTT
 10 2051 TAAAAATAAA TTGAGGCAAA CAGTTAAGTT TTATTTTATAG AGCAACAAGT
 2101 TAACTGTAAA TATTTTAATG TTAGTTTGCT CATCTATGAT CTGAGATCAT
 2151 GCCGAAGTGA GAAAAATCTC CCAAAAATAC AATTTAATGC ATTGGGAAAA
 2201 AAAAACTTTA ACAGTAATTC CAGCCACAAT CTTTAGATCA CCCTTGTAAT
 2251 GTGTTACGGG TCCATTTTTT CTGGAATCGT TTAATCTAAA GCAGTTTCCC
 15 2301 CTGTTTTGGA GATTTTGTAG TTAATTTTAA TTTTGGCTAT TGTGTGGAAA
 2351 AGATGAGCTG TCTGTGTAGA TATGAAGTAT AGTTTTTTCC ATAAAACAGA
 2401 TGTTTATTTT GTATTAATAA ATACCACTGT ACTTGTTTTA CACCATTTGT
 2451 ATACATGTGG TGATATTAAT GCTAAACTGT AAAATTCAGG AATTAATAATG
 2501 TGACCCTGTA ATTCCAAAAA AAAAAAAA

BLAST Results

25 Entry AF016448_8 from database TREMBL:
 gene: "F41EB.3"; Caenorhabditis elegans cosmid F41EB.
 Score = 390, P = 5.0e-32, identities = 73/184, positives =
 118/184,
 frame +3
 30 Entry HS211256 from database EMBL:
 human STS SHGC-15844.
 Score = 977, P = 5.5e-38, identities = 199/202

Medline entries

40 No Medline entry

Peptide information for frame 3

45 ORF from 132 bp to 980 bp; peptide length: 283
 Category: putative protein
 Classification: no clue
 50 1 MNKDLLRRVL VLMNSKHTFL ALCALRFMR R IIGLKDEFYN RYITKGNLFE
 51 PVINALLDNG TRYNNLSAV IELFEFIRVE DIKSLTAHIV ENFYKALESI
 101 EYVQTFKGLK TKYEQEKDRQ NQKLNSVPSI LRSNRRFRDA KALEEDEEMW
 151 FNEDEEEEGK AVVAPVEKPK PEDDFPDNYE KFMETKKAKE SEDKENLPKR
 55 201 TSPGGFKFTF SHSASAANGT NSKSVAQIP PATSNGSSSK TTNLPTSVTA
 251 TKGSLVGLVD YPDDEEEDDE EESSPRKRPR LGS

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_12i1, frame 3

No Alert BLASTP hits found

Pedant information for DKFZphamy2_12i1, frame 3

Report for DKFZphamy2_12i1.3

```
15  [LENGTH]  326
    [MW]      37261.10
    [pI]      5.60
    [HOMOL]   TREMBL:AF016448_8 gene: "F41E6.3"; Caenorhabditis
20  elegans cosmid F41E6. 1e-36
    [FUNCAT]  01.05.04 regulation of carbohydrate utilization [S.
    cerevisiae, YNL201c] 2e-08
    [BLOCKS]  BL00357 Histone H2B proteins
    [BLOCKS]  BP02232B
25  [BLOCKS]  PR01073C
    [BLOCKS]  BP03050C
    [BLOCKS]  BP03580F
    [BLOCKS]  PR00893F
    [KW]      All_Alpha
30  [KW]      LOW_COMPLEXITY 10.43 %

    SEQ  IVGSNKNNTICPDNYQTAQLLALILELLTFCVEHHTYHIKNYIMNKDLLRRVLVLMNSKH
    SEG  .....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
35  PRD  cccccccccccccchhhhhhhhhhhhhhhhhhhhhccchhhhhhhhhhhhhhhhhhhhhccch

    SEQ  TFLALCALRFMRRIIGLKDEFYNRYITKGNLFEPVINALLDNGTRYNLLNSAVIELFEFI
    SEG  .....
40  PRD  hhhhhhhhhhhhhhhhhhhccchhhhhccccccccchhhhhhhhhccccccccchhhhhhhhh

    SEQ  RVEDIKSLTAHIVENFYKALESIEYVQTFKGLKTKYEQEKDRQNKQLNSVPSILRSNRFR
    SEG  .....
    PRD  hheeehhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccchhhhhhhhhccccccccccccccccchhh

45  SEQ  RDAKALEEDEEMWFNEDEEEEGKAVVAPVEKPKPEDDFPDNYEKFMETKKAKESEDKENL
    SEG  .....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
    PRD  hhhhhhhhhhhhhhhhhhhccccccccccccccccccccccccchhhhhhhhhhhhhhhhhhhcc

50  SEQ  PKRTSPGGFKFTFSHSASAANGTNSKSVVAQIPPATSNNGSSSKTTNLPTSVTATKGSLVG
    SEG  .....
    PRD  cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

    SEQ  LVDYPDDEEEDEEEESSPRKRPRLG
55  SEG  .....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
    PRD  eccccccccchhhhhcccccccccccccccccccccccccccccccccccccccc
```

(No Prosite data available for DKFZphamy2_12i1.3)

(No Pfam data available for DKFZphamy2_12i1.3)

DKFZphamy2_13g19

5 group: amygdala derived

DKFZphamy2_13g19 encodes a novel 281 amino acid protein without similarity to known proteins.

10 The novel protein contains a PROSITE ASP_PROTEASE motif and seem to be expressed Ubiquitously.
No informative BLAST results; No predictive prosite, pfam or SCOP motife.

15 The new protein can find application in studying the expression profile of amygdala-specific genes.

20 unknown protein

perhaps complete cds.
Pedant: SIGNAL_PEPTIDE

25 Sequenced by EMBL

Locus: /chromosome="12p13.3"

Insert length: 2754 bp

30 Poly A stretch at pos. 2743, polyadenylation signal at pos. 2724

```

      1 GCAATCTCGG GAAATTGGAG ACTGACGCGG CTGCTCCTGC ATGTTATTTA
      51 TTTTTCCTCT TTCCCTCCCG TGGAGACCCCT CCTGTTGGAA AGAGAGCTGC
35    101 AGCACGGGAC AGAGACAGGC AGGAAGAAGC AGAGAGGACT CGGTGACGCC
      151 CCCACCGAGC AGCCCTTGGC CCACTCCTCC AGCAGGGGCC ATGAGCACCA
      201 AGCAGGAGGC CAGGAGAGAT GAGGGAGAAG CCAGGACGAG GGGGCAGGAG
      251 GCACAGCTTC GAGACCGAGC CCACCTGAGC CAGCAGCGCC GGCTCAAACA
      301 GGCCACCCAG TTCCTGCACA AGGACTCGGC CGACCTGCTC CCGCTGGACA
40    351 GCCTCAAGAG GCTCGGCACC TCCAAGGACT TGCAGCCGCG CAGTGTGATC
      401 CAAAGACGCC TGGTGGAGGG AAACCCGAAT TGGCTTCAGG GGGAGCCTCC
      451 CCGGATGCAG GACCTGATTC ATGGCCAGGA GAGCAGGAGG AAGACCAGCA
      501 GGACAGAGAT TCCAGCTCTT CTGGTCAACT GCAAGTGCCA GGACCAGCTG
      551 CTTAGAGTGG CCGTTGACAC AGGCACCCAA TACAATCGGA TCTCTGCTGG
45    601 ATGTCTCAGC CGCCTGGGGT TAGAGAAAAG GGTCTTAAAA GCCTCAGCTG
      651 GGGACCTGGC CCCTGGGGCC CCAACCCAGG TGGAGCAGTT GGAGCTACAG
      701 CTGGGGCAGG AGACTGTGGT GTGCTCGGCA CAGGTGGTGG ATGCTGAGAG
      751 TCCTGAATTC TGCCTGGGCC TGCAGACTCT GCTTTCTCTC AAGTGCTGCA
      801 TCGACCTGGA GCACGGAGTG CTGCGGCTGA AAGCCCCGTT CTCAGAGCTA
50    851 CCCTTCCTGC CTTTGTACCA AGAGCCTGGC CAGTGA CTGC TGTCTCAGTC
      901 AGTCCCCAGA GGGAAAGACC TTGCCTTAGA AGAAGAGGCG TGTGGGGAAC
      951 GGGGGCTCTT GAAGCCAGGT AGCTGGGGAC TATGGTGTCT GCCCTTCCAA
     1001 TCACCTCCCT GACCCCTGCT GTCCCATTTT CCCCAGCTGG CCGCATTCCT
     1051 CTCTGCTTCT CAGCAGCTGT CCTACTCCCC AGGACGAGTT TTTACTAGAG
55    1101 GGCCACGAT GCCAGGATTC TGATTTCATCT TCCTCCCAAG AAAAGCAAAG
     1151 CCAAATCAAG ACCACAGATA GGAACCTAAG CACAATGGGG TGCCTGCTTG
     1201 GGCTGGGTCG AAGGCTCTGC TGACTGCTGT CTTGTCCAT CACCCAATAC
     1251 CACCCCAAAC ACAACTCAAC TTCCACACC ACCATGTCTC TCACCACACC

```

1301 TTCTGGGCCT CATTATCTCC CACAACCTAGA CCGCCATGCC TCACCAACCT
 1351 ATGTCCCTGG ACCTCCTGGT GTCTGCCTCT CGGAGTCTGT GCACATCTGC
 1401 TCACAGTTGA GTGGGGGAAG AAACAGCCAG AATTCAATAC AACAAAGAGC
 1451 GGGAGTTAGT ATAGGAATGT CCATCTCATA AGGCTGAGAG CTATTTTTTC
 5 1501 CTGTGGCTGC AAATGTCTGA AGCCAGTTAG TTTGATTACC CTGTGCAAAA
 1551 CCTTGGACAT ACTTCTGCTA TTAACGCTAT AGGTATTTAT CCGTTTCCAC
 1601 TGGCTTTTTG TACCCACCGA GCCCTGAGC CTTGCGTGTG TGTGTGTGGA
 1651 AGAGCCTTGT AGAGAACTGC TCCTGTGAGG CAGACAGGAC AGTGAGGTTG
 1701 TCACCACTCA GACTTCACCT ATTCAGCATT CTTTCTGATT TCTAGAACTA
 10 1751 TCCACCTCAT TAGGCCTTCT TCCTATCCCC ATCTCTGGCC TCTTGAGCTT
 1801 AAGCTTGTAT TGTCTTGGAA TCAGTGGCTT TCTAACCCCC TGCCAGGCTT
 1851 TGCCAAAGCA AAAAGACAGA GGCTTTTTTT TTTTTTTTAA AGTTTGGGGT
 1901 CTGTCAGGAG ACAGAGGCTT TTTTGAATTC ACTGTGAAGA GAAGAACCCG
 1951 AACCTTAAGA CGCCAGATCC CTGAGAGTCT TTCTGGCTGG TTTGAGTCTC
 15 2001 TCAAATCATG GATTAGGAGT AAAGAAAGAG GCAGGCGCAA TGGCTCATGC
 2051 CTGTAATCCC AGCACTTTGG GAGGCTGAGG TGGGTGGATC ACTTGAGGTC
 2101 AGGAGTTTGA GACCAGCCTG GGTAAATATGG CAAAACCCCA TCTCTACTAA
 2151 AAAATACAAA AATTAGCCAG GTATGGTGGT GAACACCTGT AATCCCAGCT
 2201 ACTTGGGAAGG CTGAGGCATA GGAGTTGCTT GAACCTGGGA GATGGGGGTT
 20 2251 GTAGTGAGCC AAGTTCGTGC CATCGGACTC CAGCCTGGGT GAAGGAGTGA
 2301 GACCCTGTCT CCAAAAACAA ACAAAAAGG AGCAGAGAAA GACAGTGGTA
 2351 CAGCTAACCT GAACAAGGGA ACTGGGACCG TTGGGCTGAA ACAGTCTTGA
 2401 GCCTGGGGTT GACTGGGTGA GAGAAGAACC GGGATGCAAG GAGCTGCCTG
 2451 TGACACCTGG CCTGCCCTTT CTCAGCTGCC TCCCCTGCC TTTCTCAGCT
 25 2501 GCCTCCCCTG CCCTCAGAAG GAAAGGAGAG GGCTCACTTA TCACTTGTGC
 2551 CATAGCACCT GGTCTCAAAA TCCTAAAAGC TTTCTCGCC CTCCTGCCT
 2601 TGCTCCACAA GGTCCACTTT CCTGGGTCTT GTGCTGTGCC TTTCTTGTG
 2651 TGCCTCCTGC TGCTTCTGTA ACTGCAGACC CCAGGCCCAA TTGCAAGCCC
 2701 TCGGCTCAGC TGCTTCTCCA TTGGAATAAA CTCTTGTTTC TCTAAAAAAA
 30 2751 AAAA

BLAST Results

35

No BLAST result

Medline entries

40

No Medline entry

45

Peptide information for frame 2

50 ORF from 41 bp to 883 bp; peptide length: 281
 Category: putative protein
 Classification: no clue
 Prosite motifs: ASP_PROTEASE (173-184)

55

1 MLFIFPLSLP WRPSCWKESC STGQRQAGRS REDSVTPPPS SPWPTPPAGA
 51 MSTKQEARRD EGEARTGQE AQLRDRAHLS QQRRLKQATQ FLHKDSADLL
 101 PLDSLKRLGT SKDLQPRSVI QRRLVEGNPN WLQGEPPRMQ DLIHQESRR
 151 KTSRTEIPAL LVNCKCQDQL LRVAVDTGTQ YNRISAGCLS RLGLEKRVLK

201 ASAGDLAPGP PTQVEQLELQ LGQETVVCSA QVVDAESPEF CLGLQTLLSL
 251 KCCIDLEHGV LRLKAPFSEL PFLPLYQEPG Q

5

BLASTP hits

No BLASTP hits available

10

Alert BLASTP hits for DKFZphamy2_13g19, frame 2

PIR:S50646 hypothetical protein YER143w - yeast (Saccharomyces
 cerevisiae), N = 1, Score = 90, P = 0.26

15

TREMBL:RND060_1 product: "DNA (cytosine-5-)-methyltransferase";
 Rattus
 norvegicus mRNA for DNA (cytosine-5-) -methyltransferase, partial
 cds.,
 N = 1, Score = 81, P = 0.89

20

>PIR:S50646 hypothetical protein YER143w - yeast (Saccharomyces
 cerevisiae)

Length = 428

25

HSPs:

Score = 90 (13.5 bits), Expect = 3.0e-01, P = 2.6e-01
 Identities = 28/112 (25%), Positives = 48/112 (42%)

30

Query: 155 TEIPALLVNCKCQDQLLRVAVDGTGTQYNRISAGCLSRLGLEKRVLKASAGD-
 --LAPGPP 211
 T++P L +N + + ++ VDTG Q +S + GL + + K G+

+ G

35

Sbjct: 199
 TQVPMLYINIEINNPVKAFVDGTGAQTTIMSTRLAKKTGLSRMIDKRFIDGEARGVGTGKI 258

Query: 212 XXXXXXXXXXXXXXXXXXXX-
 CSAQVVDAESPEFCLGLQTLLSLKCCIDLEHGVRL 263

40

CS V+D + + +GL L C+DL+

VLR+

Sbjct: 259 IGRHQAAQVKIETQYIPCSFTVLDTDI-
 DVLIGLDMLKRHLACVDLKENVLRI 310

45

Pedant information for DKFZphamy2_13g19, frame 2

Report for DKFZphamy2_13g19.2

50

[LENGTH] 281
 [MW] 31330.97
 [pI] 8.75
 55 [BLOCKS] PRO0049D
 [BLOCKS] BP019216
 [PROSITE] ASP_PROTEASE 1
 [KW] All_Alpha

```

[OKW]          LOW_COMPLEXITY          9.96 %

```

```

SEQ  CLGLQTLTSLKCCIDLEHGVLRLKAPFSELPFLPLYQEPGQ
SEG  .....
PRD  cccchhhhhhhhhhhcchhhhhhhcccccccccccccccccc

```

Prosites for DKFZphamy2_13g19.2

PS00141 173->185 ASP_PROTEASE PDOC00128

(No Pfam data available for DKFZphamy2_13g19.2)

DKFZphamy2_14b5

5 group: intracellular transport and trafficking

DKFZphamy2_14b5 encodes a novel 771 amino acid protein which shows 61% identity to the human TYL protein and 48% identity to the human Tic protein.

10

Both proteins show similarity to Sec7 of *Saccharomyces cerevisiae*, which takes function in vesicular trafficking. The new protein shows also significant similarity to human ARN03, which is involved in the control of Golgi structure and function.

15

DKFZphamy2_14b5 is predominantly expressed in the CNS and germ cells.

20

The new protein can find application in diagnosis/therapy of diseases related to vesicular trafficking e.g. in synapses of the central nervous system and in studying expression profiles.

similarity to TYL protein (Homo sapiens)

25

Sequenced by EMBL

Locus: /map="445.7 cR from top of Chr5 linkage group"

30

Insert length: 4528 bp

Poly A stretch at pos. 4511, polyadenylation signal at pos. 4489

```

1 CTCGCTCAGC CTCTCCACAT CGCGGCTCCG GCACCTGAAG GGACGCGGGC
35 51 GGGCGCGGGC AGCTCCGACC GGGCGGCGCG GGGCGGGACA GGCAGCCCCG
101 CGGCCTCCG/ TGGCCCCGCC GTGAGAGGCC GGACCCGCGG CGGGGACCAG
151 CAGCGGTCT/ AGGAGTCCC AGGAGCAGCC AGGACAGGCG GAAGCAGTGG
201 CTGCCATGG/ :AGGACAAG CTCTTATCTG CAGTGCCTGA GGAAGGCGAT
251 GCCACCCGT/ :CCCCGGTCC AGAGCCTGAA GAGGAGCCAG GGGTCCGGAA
40 301 TGGGATGGC :GTAGGGGCC TGAACAGCAG CCTCTGCAGC CCAGGGCACG
351 AGCGAAGGG/ :ACCCAGCG GACACTGAGG AACCCACGAA GGACCCAGAT
401 GTGGCCTTCC :TGGCCTCAG CCTTGGCCTC TCTCTACCA ATGGCCTAGC
451 CCTGGGGCCA :ACTTGAACA TTCTGGAAGA TTCAGCGGAG TCCAGGCCCT
501 GGAAGGGCTGG CGTGCTGGCA GAGGGGGACA ATGCTTCCAG GAGCCTCTAC
45 551 CCAGATGCTG AGGACCCCTCA GCTGGGGTTG GATGGTCCCG GGGAGCCAGA
601 TGTGCGGGAT GGCTTCAGCG CCACGTTTGA GAAGATTCTG GAGTCAGAGC
651 TGCTGCGGGG CACCCAGTAC AGCAGCCTCG ACTCCCTAGA CGGGCTGAGC
701 CTCACGGATG AGAGCGACAG CTGCGTCAGC TTCGAGGCCC CCTCACACC
751 CCTCATCCAG CAGCGGGCCC GTGACAGCCC TGAGCCAGGG GCTGGGTTGG
50 801 GCATTGGGGA CATGGCGTTT GAGGGGGACA TGGGGGCAGC TGGTGGTGAT
851 GGGGAGCTGG GCAGCCCCCT GCGGCGCTCC ATCTCCAGCA GCCGCTCTGA
901 GAATGTCCTG AGCCGCTGT CTCTCATGGC CATGCCCAAT GGATTCCATG
951 AAGATGGCCC TCAGGGCCCA GGGGGGGATG AGGATGATGA TGAGGAGGAC
1001 ACGGACAAGT TGCTGAACTC AGCCAGTGAC CCCAGCCTGA AGGATGGCCT
55 1051 GTCAGACTCA GACTCTGAGC TCAGCAGCTC GGAGGGGTTG GAGCCTGGTA
1101 GTGCAGACCC TCTGGCCAAC GGGTGCCAGG GGGTCAGTGA AGCTGCTCAT
1151 CGGCTGGCAC GCCGTCTCTA CCACCTCGAG GGCTTCCAGC GCTGTGATGT
1201 GGGCCGGCAG CTGGGCAAGA ACAACGAGTT TAGCAGGCTG GTGGCCGGGG

```

	1251	AGTACCTCAG	TTTCTTCGAC	TTCTCGGGCT	TGACTCTGGA	CGGAGCACTC
	1301	AGAACATTCT	TGAAGGCCCT	CCCCGCTGATG	GGGGAGACAC	AAGAGCGTGA
	1351	GCGGGTCCTC	ACACACTTCT	CCCCCGGTA	CTGCCAGTGC	AACCCTGATG
	1401	ACAGCACTTC	GGAAGATGGG	ATCCACACGC	TCACCTGTGC	CCTGATGCTG
5	1451	CTCAACACGG	ACCTGCACGG	CCACAACATT	GGCAAAAAGA	TGTCTGTCA
	1501	GCAATTCATT	GCCAACCTTG	ACCAGCTGAA	TGATGGCCAA	GACTTTGCCA
	1551	AAGACCTGCT	GAAGACCTT	TACAACCTCA	TCAAGAATGA	AAAGCTGGAA
	1601	TGGGCCATTG	ATGAGGATGA	GCTGAGGAAA	TCCCTGTCTG	AGCTGGTGGG
	1651	TGACAAGTTC	GGGACAGGCA	CGAAGAAGGT	GACGCGAATC	CTGGATGGTG
10	1701	GCAACCCCTT	CCTGGATGTC	CCACAGGCGC	TCAGTGCCAC	CACCTACAAG
	1751	CACGGCGTCC	TGACCCGGAA	GACTCACGCT	GACATGGATG	GCAAGAGGAC
	1801	GCCCCGTGGG	AGGCGTGGCT	GGAAGAAATT	CTACGCAGTG	CTCAAAGGGA
	1851	CCATCCTGTA	CCTGCAGAAG	GATGAGTACA	GGCCTGACAA	AGCTCTATCG
	1901	GAGGGTGACC	TGAAGAACGC	CATTGCGGTG	CATCACGCTC	TGGCCACCA
15	1951	GGCCTCTGAC	TACAGCAAGA	AGTCCAACGT	GCTGAAGCTT	AAGACAGCCG
	2001	ACTGGAGGGT	ATTCCTCTTC	CAGGCACCGA	GCAAGGAAGA	AATGCTGTCC
	2051	TGGATCCTCA	GGATCAACCT	GGTGGCAGCC	ATCTTCTCTG	CCCCGGCCTT
	2101	CCCAGCCGCT	GTCAGCTCCA	TGAAGAAGTT	CTGTCGGCCC	CTGCTGCCCT
	2151	CCTGCACCAC	CCGCTCTGCT	CAGGAGGAGC	AACTGCGGTC	TCATGAGAAT
20	2201	AAGTTGAGGC	AGCTGACTGC	GGAGCTGGCC	GAACACAGGT	GTACCCAGT
	2251	CGAGAGGGGC	ATCAAGTCCA	AGGAGGCCGA	GGAGTACCGG	TTGAAGGAGC
	2301	ACTATCTCAC	CTTCGAGAAA	AGCCGTTATG	AGACCTATAT	CCACCTCTG
	2351	GCTATGAAAA	TCAAAGTGGG	CTCAGATGAT	CTGGAGCGGA	TTGAGGCCCG
	2401	GCTGGCCACT	CTGGAAGGGG	ATGACCCTTC	TCTCCGGAAG	ACACATTCAA
25	2451	GCCCTGCCCT	CAGCCAGGGC	CATGTGACTG	GCAGCAAAAC	CACAAAGGAT
	2501	GCCACTGGGC	CTGATACTTA	GCTGACATGG	ATTTGCAGAC	CCCAGGGTGG
	2551	GCAGATGTCT	CCAGTGGGGT	CAGTGAGCAC	AATTCCAGCC	AGGGGCCACT
	2601	TGGACCAAGC	TCCAGTCAGT	TGATGGGCAG	CTAGAGGGGT	GCAGAAAGCC
	2651	TGTGGGCCCC	GGAGATGGAG	ATGCCGTTTG	TGGCGTTGAT	CTCCTTGCGT
30	2701	CCTTGGGCA	CTCCGGGCAT	CAGACCCTCT	CCCTGGCCCT	TGTTTTCTC
	2751	TCCACCATGG	AGCCTCATTT	TGTAGGCCAG	TTGTGTGCAT	GCTCTAGACA
	2801	CCACCTCGCT	GGAGAAGCTG	GAAGGGCTGT	TGTCTTCCCA	GGTCTTTCTC
	2851	TTCTCATCAA	GCTCCTCTCC	TCATCTTTT	TGTGTGTGAG	GGCAGGTCTT
	2901	GACTCTAGGT	CTCAGCTGGA	ACCCACCCCT	TTCTCCTCCT	CCTTCTCTG
35	2951	AGTTGACCAG	CAGCAGGTCT	GCCGACCACC	AGCACCATCC	TCTCCTCCCA
	3001	GCAGCCTCCA	GAACCATGCC	CAGGTCTCCT	GCCTCACATC	ACAATAATCT
	3051	GGGACCCAGG	CTTGTCCTCT	TTGAGTGTAA	AGCTGACTCC	ATCATATGTG
	3101	CATCCACTTC	TTTTTCATCCA	TTGAGATCAC	ACTGCCTCCT	TTTTATACAG
	3151	ACACAAATAT	ACATCTATAA	GAATAATATA	TACATAAGGA	ACCCCTGAAA
40	3201	GATGGTTTTG	GAACCTGGAAT	CAGTTAGAGG	ATGAAATCAG	ATAAAGGAAA
	3251	AGCCTATTTT	GGAGCTTCCC	CTGTTAGGAA	GGATGGCTGC	ACCTGGCCCC
	3301	CTGGCATTCC	TGACGCTCTA	GGAGGGAAGG	GGGAGGCAGT	GCTGGCCTCC
	3351	CTTGCCCTGT	TTTTCCCTCT	TCCAGCTGAC	CTGTGACTTA	TACTGCTCTT
	3401	ACCGATGATA	CTTTTGGAAA	AAATAGAGCG	TGTATGCACC	GCCCCGTTTTG
45	3451	TCCCATGGAT	ATCCTGGGGT	GTGAGTCGGA	TGGGACCACG	GCCCTGTTTA
	3501	TATTGGGGTC	TTTATGTTGG	TGCTGCCAGG	TCTCTGAGCT	CCAGAGGTGG
	3551	CCTCTTGGAC	AGATCTACTG	CTATAGGAAT	AAAAGACACT	CTGTCTCGCA
	3601	AATGGCTGCT	TGTCAACAAG	CCCAAAGATG	CTTGTCGGAG	GACGGTTATG
	3651	GAAGCCCTTA	ATTCTTGGTT	GTGGGAAAAG	GTGGAATGAC	AAGTTATTGA
50	3701	TTGTTTTTCT	GTCGCTATTT	CTTTCATTTG	TCTAGTGAAT	CAGAAAGGCT
	3751	TAGCCAAGGC	CACATCTGGG	AAGAGTGGAG	AAATTTGCCA	CTTGACGATC
	3801	ACGGATTAGC	TAGCACCTTT	AAGCCCTGCA	TTTCTCCAAC	TGACAAGTGG
	3851	GTGGGGGTGA	TGGCACATTC	AGTGTGGCTA	TGAAGAGCGA	ATCCTCTCTA
	3901	TTGTTTAAAT	AGATTACTGT	AGTTTGGCCA	GGAATTTGGC	GTCAGTGGTA
55	3951	ACACACTTAG	TAAATAAAT	AAGCCAGGCT	TGCAACTAAG	TATCTAAGTT
	4001	TACAGGCCCA	CTCACATTTG	AGGCAAGGGG	CTATTGAGTA	TGTGGAGAGT
	4051	TGTAAGTATT	TAAATTCAGA	TTATTTAAGT	TGGATCAGCT	GAAGTGTGTT
	4101	TTAGACCCAA	ACCATCTGGC	CCCTTCGTTT	TGCTCAGAGG	AAGTAAATGT

```

4151 TCACTTAAAT GAAATTGAAA ACGCCATGTG GCACCACAAA AGAGCTCTCT
4201 GTACTTTCCC CATGCTGCCT CAAAAGTTCT GTGAGTTTCG GGGTCAGTGT
4251 CCCACCCCTC ACTTCCCGAG GGC GG GTGAG TGGAGAGCAG AGCCAGGAGC
4301 TCTGGCAGCT GTGGACAGAT GTGCTTCCTG AGCATGGGTT GTGCCTCCCA
5 4351 TCAGTAAAAA AATGTTTAGT TCACTTCCTT AATTGTATAA TTATTTATTT
4401 GTAAATTATA TACATGTACT ACTGTACTAA AATATTATGT ACATTATAAA
4451 ACATACACAA AAATAGAAAT TTAATAAAGA TGAGATGAAA ATAAATCTAA
4501 GTCAAAGTTC CAAAAAATAA AAAAAAAA

```

BLAST Results

No BLAST result

Medline entries

98086482:
 Perletti L, Talarico D, Trecca D, Ronchetti D, Fracchiolla NS,
 Maiolo AT, Neri A.; Identification of a novel gene, PSD, adjacent
 to
 NFKB2/1yt-10, which contains Sec7 and pleckstrin-homology
 domains.
 Genomics 46:251-259(1997)

Peptide information for frame 2

ORF from 206 bp to 2518 bp; peptide length: 771
 Category: similarity to known protein
 Classification: Cell signaling/communication

```

1 MEEDKLLSAV PEEGDATRDP GPEPEEPEGV RNMASEGLN SSLCSPGHER
51 RGTPADTEEP TKDPDVAFHG LSLGLSLTNG LALGPDNLIL EDSAESRPWR
40 101 AGVLAEGDNA SRSLYPDAED PQLGLDGPGE PDVRDGFSAF FEKILESELL
151 RGTQYSSLDL LDGLSLTDES DSCVSFEAPL TPLIQQRARD SPEPGAGLGI
201 GDMAFEGDMG AAGGDGELGS PLRRSISSSR SENVLSRLSL MAMPNGFHED
251 GPQGPGGDED DDEEDTDKLL NSASDPSLKD GLSDSDSELS SSEGLEPGSA
301 DPLANGCQGV SEAAHRLARR LYHLEGFQRC DVARQLGKNN EFSRLVAGEY
45 351 LSFFDFSGLT LDGALRTFLK AFPLMGETQE RERVLTHFSR RYCQCNPDSS
401 TSEDGIHTLT CALMLLNTDL HGHNIGKKMS CQQFIANLDQ LNDGQDFAKD
451 LLKTLYNSIK NEKLEWAIDE DELRKSLSEL VDDKFGTGTK KVTRILDGGN
501 PFLDVPQALS ATTYKHGVLK RKTHADMDGK RTPRGRRGWK KFYAVLKGTI
551 LYLQKDEYRP DKALSEGDLK NAIRVHHALA TRADYSKKS NVLKLKTADW
50 601 RVFLFQAPSK EEMLSWILRI NLVAAIFSAP AFPAAVSSMK KFCRPLLPSK
651 TTRLCQEEQL RSHENKLRQL TAELAHRCH PVERGIKSKE AEEYRLKEHY
701 LTFEKSRYET YIHLLAMKIK VGSDDLRIE ARLATLEGDD PSLRKTHSSP
751 ALSQGHVTGS KTTKDATGPD T

```

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_14b5, frame 2

5 PIR:G01205 TYL protein - human, N = 2, Score = 1421, P = 8.6e-150

TREMBL:AB023159_1 gene: "KIAA0942"; product: "KIAA0942 protein";
Homo
sapiens mRNA for KIAA0942 protein, partial cds., N = 1, Score =
10 1251, P
= 2.3e-127

TREMBL:U63127_1 gene: "TIC"; product: "Tic"; Human SEC7 homolog
Tic
15 (TIC) mRNA, complete cds., N = 1, Score = 1050, P = 4.6e-106

>PIR:G01205 TYL protein - human
Length = 645

20 HSPs:

Score = 1421 (213.2 bits), Expect = 8.6e-150, Sum P(2) = 8.6e-
150
25 Identities = 280/452 (61%), Positives = 336/452 (74%)

Query: 301
DPLANGCQGVSEAAHRLARRLYHLEGFQRCQDVARQLGKNNFEFSRLVAGEYLSFFDFSGLT 360
D L+NG + EAA RLA+RLY L+GF++ DVAR LGKNN+FS+LVAGEYL
30 FF F+G+T
Sbjct: 166
DTLSNGQKADLEAAQRLAKRLYRLDGFRKADVARHLGKNNDFSKLVAGEYLKFFVFTGMT 225

Query: 361
35 LDGALRTFLKAFPLMGETQERERVLTHFSRRYCQCNPDDSTSEDGIHTLTALMMLNTDL 420
LD ALR FLK LMGETQERERVL HFS+RY QCNP+ +SEDG
HTLTALMMLNTDL
Sbjct: 226
LDQALRVFLKELALMGETQERERVLAHFSQRYFQCNPPEALSSSEDGAHTLTALMMLNTDL 285

Query: 421
40 HGHNIGKKMSCQQFIANLDQLNDGQDFAKDLLKTLYNSIKNEKLEWAIDEDELRSLSSEL 480
HGHNIGK+M+C FI NL+ LNDG DF ++LLK
LY+SIKNEKL+WAIDE+ELR+SLSEL
45 Sbjct: 286
HGHNIGKRMTCGDFIGNLEGLNDGGDFPRELLKALYSSIKNEKLQWAIDEEELRRSLSEL 345

Query: 481 VDDKFGTGTKKVTRIL----
50 DGGNPFLDVPQALSATTYKHGVLTRKTHADMDGKRTPRGR 536
D K + RI G +PFLD+ A YKHG L RK HAD D
++TPRG+
Sbjct: 346 ADPN----
PKVIKRISGGSGSGSSPFLDLTPEPGAAYKHGALVRKVHADPDCRKTPRGK 401

55 Query: 537
RGWKKFYAVLKGITILYLQKDEYRPDKALSEGDLKNAIRVHHALATRASDYSSKSNVLKLLK 596
RGWK F+ +LKG ILYLQK+EY+P KALSE +LKNAI
+HHALATRASDYSK+ +V L+

Sbjct: 402
RGWKS F H G I L K G M I L Y L Q K E E Y K P G K A L S E T E L K N A I S I H H A L A T R A S D Y S K R P H V F Y L R 461

Query: 597
5 TADWRVFLFQAPSKEEMLSWILRINLXXXXXXXXXXXXXXXXSMKKFCRPLLPSCTTRLCA 656
TADWRVFLFQAPS E+M SWI RIN+ S KKF

RPLLPS TRL Q
Sbjct: 462
TADWRVFLFQAPSLQMQSWITRINVVAAMFSAPPFPAAVSSQKKFSRPLLPSAATRLSQ 521

Query: 657
EEQLRSHENKLRQLTAELAEHRCHPVERGIKSKEAEEYRLKEHYLTFEKSRYETYIHLA 716
EEQ+R+HE KL+ + +EL EHR + + + KEAEE R KE YL FEKSRY
TY LL

Sbjct: 522
EEQVRTHEAKLKAMASELREHRAAQLGKKGRGKEAEEQRQKEAYLEFEKSRYSTYAALLR 581

Query: 717 MKIKVGSDDLRIEARLATLEGDDPSLRKTHSSPAL 752
+K+K GS++L+ +EA LA + L +HSSP+L

Sbjct: 582 VKLKAGSEELDVAEALAQAGSTEDGLPPSHSSPSL 617

Score = 63 (9.5 bits), Expect = 8.6e-150, Sum P(2) = 8.6e-150
Identities = 19/64 (29%), Positives = 23/64 (35%)

Query: 132 DVRDGFSAFTEKILESELLRGTQYXXXXXXXXXXXXXXXXXXXX-
CVSFEAPLTPLIQQRARD 190
D D FS FE ILES +GT Y +FE P P

Sbjct: 18
30 DGPDSFSCVFEAILESHRAKGTSYTSLASLEALASPGPTQSPFFTFELPPQPPAPRPDPP 77

Query: 191 SPEP 194
+P P

Sbjct: 78 APAP 81

Pedant information for DKFZphamy2_14b5, frame 2

Report for DKFZphamy2_14b5.2

45 [LENGTH] 771
[MW] 84660.55
[PI] 5.04
[HOMOL] PIR:G01205 TYL protein - human 1e-158
[FUNCAT] 30.09 organization of intracellular transport vesicles
[S. cerevisiae, YDR170c] 5e-22
50 [FUNCAT] 30.08 organization of golgi [S. cerevisiae, YDR170c]
5e-22
[FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae,
YDR170c] 5e-22
[FUNCAT] 08.07 vesicular transport (golgi network, etc.) [S.
cerevisiae, YDR170c] 5e-22
55 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YPR095c]
4e-04
[BLOCKS] BL012778
[BLOCKS] BP02373F

-101-


```

5  SEQ  VDDKFGTGTKKVTRILDGGNPFLDVPQALSATTYKHGVLTRKTHADMDGKRTPRGRRGWK
   SEG  .....
   lbtn- .....EEEEEEEEETTEET--
   TTTCEE

10  SEQ  KFYAVLKGTILYLQKDEYRPDKALSEGDLKNAIRVHHALATRASDYSKKS NVLKLKTADW
   SEG  .....
   lbtn- EEEEEETTEEEECCHHHHHHCCBTTT-
   TCCEETTTTEEEETTTTCTTTEEEETTTT

15  SEQ  RVFLFQAPSKEEMLS WILRINLVAAIFSAPAFPAAVSSMKKFCRPLLPSCTTRL CQEEQL
   SEG  .....XXXXXXXXXXXXXXXX.....
   lbtn- CEEEEECCHHHHHHHHHHH.....

20  SEQ  RSHENKLRQLTAELAEHRCHPVERGIKSKEAEEYRLKEHYLTFEKSRYETIYHLLAMKIK
   SEG  .....
   lbtn- .....

25  SEQ  VGSDDLIERIEARLATLEGDDPSLRKTHSSPALSQGHVTGSKTTKDATGPD
   SEG  .....
   lbtn- .....

```

(No Prosite data available for DKFZphamy2_14b5.2)

30 Pfam for DKFZphamy2_14b5.2

HMM_NAME PH (pleckstrin homology) domain

```

35  HMM
   *dvIREGWMyKWgswrkstg.....nWqrRWFvLrndpnrLiYYkddk
                                   + ++G + +++ + ++          W++ ++VL++ + L++
   KD+
   Query           512  TTYKHGVLTRKTHADMDGKRTPRGRRGWKKFYAVLKG--
40  TILYLQKDE-      557

   HMM
   dekPr.....YMIIdld.cW rMidVEidWmmdndHCFiIWtrq.rtyYF
                                   +P+          +++++ + ++D ++ +++ +++T +
45  R+++F
   Query           558  -YRPDKALSEGDLKNAIRVHHALATRASDYSKK-
   SNVLKLKTADWRVFLF    605

   HMM
   QAeNeEEMmeWMSaIrRaIw*
50  QA+++EEM +W+ I+ + +
   Query           606  QAPSKEEMLS WILRINLVAA          625

```

DKFZphamy2_14m1b

5 group: transcription factors

DKFZphamy2_14m1b.pl encodes a novel 252 amino acid protein with similarity to the homeotic protein emx2 of man, mouse and zebra fish as well as to the gene "empty spiracles" of *Drosophila melanogaster*.

Homoeobox genes are known to play important roles in developmental processes. In zebrafish emx2 mRNAs are found in the dorsal telencephalon, parts of the diencephalon and the otocyst. The human homologue Emx2 appears to be already expressed in 8.5 day embryos. It is also expressed in the presumptive cerebral cortex, olfactory bulbs, in some neuroectodermal areas in embryonic head including olfactory placodes in earlier stages and olfactory epithelia later in development. Mutants of the *D. melanogaster* gene "empty spiracles" display spiracles devoid of filzkörper, no antenna and an open head.

The new protein can find application in modulating the expression of genes controlled by this transcription factor and modulation of neuronal development.

strong similarity to homeotic protein emx2 (*Homo sapiens*)

perhaps differential splicing

Sequenced by EMBL

Locus: /chromosome="10"

Insert length: 241b bp
Poly A stretch at pos. 2398, polyadenylation signal at pos. 2373

```

1  GAAAAAAAAA GAAAAAAAAA GAAAAAAAAAT TACCCCAATC CACGCCTGCA
40  51 AATTCTTCTG GAAGGATTTT CCCCCTCTC TTCAGGTTGG GCGCGTTTGG
    101 TGCAAGATTC TCGGGATCCT CGGCTTTGCC TCTCCCTCTC CCTCCCCCT
    151 CCTTTCCTTT TTCCTTTCCT TTCCTTTCTT TCTTCCTTTC CTTCCCCCA
    201 CCCCCACCCC CACCCCAAAC AAACGAGTCC CCAATTCTCG TCCGTCTCTG
    251 CCGCGGGCAG CGGGCGGCGG AGGCAGCGTG CGGCGGTGCG CAGGAGCTGG
45  301 GAGCCCAGGG CGCCCGCTCC TCGGCGCAGC ATGTTCCAGC CGGCGCCCAA
    351 GCGCTGCTTC ACCATCGAGT CGCTGGTGGC CAAGGACAGT CCCCTGCCCCG
    401 CCTCGCGCTC CGAGGACCCC ATCCGTCCCC CGGCACTCAG CTACGCTAAC
    451 TCCAGCCCCA TAAATCCGTT CCTCAACGGC TTCCACTCGG CCGCCGCCGC
    501 CGCCGCCGGT AGGGGCGTCT ACTCCAACCC GGACTTGGTG TTCGCCGAGG
50  551 CGGTCTCGCA CCCGCCCAAC CCCGCCGTGC CAGTGCACCC GGTGCCGCCG
    601 CCGCACGCCC TGGCCGCCCA CCCCCTACCC TCCTCGCACT CGCCACACCC
    651 CCTATTCGCC TCGCAGCAGC GGGATCCGTC CACCTTCTAC CCCTGGCTCA
    701 TCCACCGCTA CCGATATCTG GGTCATCGCT TCCAAGGGAA CGACACTAGC
    751 CCCGAGAGTT TCCTTTTGCA CAACGCGCTG GCCCGAAAGC CCAAGCGGAT
55  801 CCGAACC GCC TTCTCCCCGT CCCAGCTTCT AAGGCTGGAA CACGCCTTTG
    851 AGAAGAATCA CTACGTGGTG GGCGCCGAAA GGAAGCAGCT GGCACACAGC
    901 CTCAGCCTCA CGGAAACTCA GGTAAAAGTA TGGTTTCAGA ACCGAAGAAC
    951 AAAGTTCAAA AGGCAGAAGC TGGAGGAAGA AGGCTCAGAT TCGCAACAAA

```

1001 AGAAAAAAGG GACGCACCAT ATTAACCGGT GGAGAATCGC CACCAAGCAG
 1051 GCGAGTCCGG AGGAAATAGA CGTGACCTCA GATGATTAAA AACATAAACCC
 1101 TAACCCCAACA GAAACGGACA ACATGGAGCA AAAGAGACAG GGAGAGGTGG
 1151 AGAAGGAAAA AACCTACAA AACAAAAACA AACCGCATAAC ACGTTCACCG
 5 1201 AGAAAGGGAG AGGGAATCGG AGGGAGCAGC GGAATGCGGC GAAGACTCTG
 1251 GACAGCGAGG GCACAGGGTC CCAAACCGAG GCCGCGCCAA GATGGCAGAG
 1301 GATGGAGGGCT CCTTCATCAA CAAGCGACCC TCGTCTAAAG AGGCAGCTGA
 1351 GTGAGAGACA CAGAGAGAAG GAGAAAGAGG GAGGGAGAGA GAGAAAGAGA
 1401 GAGAAAGAGA GAGAGAGAGA GAGAGAAAGC TGAACGTGCA CTCTGACAAG
 10 1451 GGGAGCTGTC AATCAAACAC CAAACCGGGG AGACAAGATG ATTGGCAGGT
 1501 ATTCCGTTTA TCACAGTCCA CTAAAAAAT GATGATGATG ATAAAAACCA
 1551 CGACCCAACC AGGCACAGGA CTTTTTTGTT TTTTGCACCT CGCTGTGTTT
 1601 CCCCCCATC TTTAAAAATA ATTAGTAATA AAAAACAAAA ATTCCATATC
 1651 TAGCCCCATC CCACACCTGT TTCAAATCCT TGAAATGCAT GTAGCAGTTG
 15 1701 TTGGGCGAAT GGTGTTTAAA GACCGAAAAAT GAATTGTAAT TTTCTTTTCC
 1751 TTTTAAAGAC AGGTTCTGTG TGCTTTTAT TTTGATTTT TTTCCCAAGA
 1801 AATGTGCAGT CTGTAAACAC TTTTGTATAC CTTCTGATGT CAAAGTGATT
 1851 GTGCAAGCTA AATGAAGTAG GCTCAGCGAT AGTGGTCCTC TTACAGAGAA
 1901 ACGGGGAGCA GGACGACGGG GGGGCTGGGG GTGGCGGGGG AGGGTGCCCA
 20 1951 CAAAAAGAAT CAGGACTTGT ACTGGGAAAA AAACCCCTAA ATTAATTATA
 2001 TTTCTTGGAC ATTCCCTTTC CTAACATCCT GAGGCTTAAA ACCCTGATGC
 2051 AAACCTTCTC TTTCACTGGT TGGAGAAATT GGCCGAGTTC AACCATTCAC
 2101 TGCAATGCCT ATTCCAACT TTAATCTAT CTATTGCAAA ACCTGAAGGA
 2151 CTGTAGTTAG CGGGGATGAT GTTAAGTGTG GCCAAGCGCA CGGCGCAAG
 25 2201 TTTTCAAGCA CTGAGTTTCT ATTCCAAGAT CATAGACTTA CTAAAGAGAG
 2251 TGACAAATGC TTCCTTAATG TCTTCTATAC CAGAATGTAA ATATTTTGT
 2301 GTTTTGTGTT AATTTGTTAG AATTCTAACA CACTATATAC TTCCAAGAAG
 2351 TATGTCAATG TCAATATTTT GTCAATAAAG ATTTATCAAT ATGCCCTCAC
 2401 AAAAAAAAAA AAAAAA

BLAST alert EMBL/EMBLNEW

35 EMBLNEW:AL133353 Human DNA sequence *** SEQUENCING IN PROGRESS
 *** from
 clone RP11-483F11; N = 2, Score = 3108, P = 5.3e-134

40 EMBL:HSEMX2 H.sapiens EMX2 mRNA; N = 1, Score = 2385, P = 5.1e-101

Medline entries

45 92331606:
 Simeone A, Gulisano M, Acampora D, Stornaiuolo A, Rambaldi M,
 Boncinelli E.;
 Two vertebrate homeobox genes related to the Drosophila empty
 spiracles gene are expressed in the embryonic cerebral cortex.
 50 EMBO J
 1992 Jul;11(7):2541-50

55

Peptide information for frame 1

ORF from 331 bp to 1086 bp; peptide length: 252
 Category: questionable ORF
 Classification: unset
 Prosite motifs: HOME0B0X_1 (187-210)

5

1 MFQAPAKRCF TIESLVAKDS PLPASRSED P IRPAALSYAN SSPINPFLNG
 51 FHSAAAAAAG RGVYSNPDLV FAEAVSHPPN PAVPVHPVPP PHALAAHPLP
 101 SSHSPHPLFA SQQRDPSTFY PWLIHRYRYL GHRFQGNDTS PESFLLHNAL
 151 ARKPKRIRTA FSPSQLLRL HAFEKNHYVV GAERKQLAHS LSLTETQVKV
 201 WFQNRRTKFK RQKLEEEGSD SQQKKKGTHH INRWRIATKQ ASPEEIDVTS
 251 DD

10

15

Alert BLASTP hits for DKFZphamy2_14m1b, frame 1

PIR:I51737 homeotic protein emx2 - zebra fish; N = 2, Score =
 753, P =
 1e-105

20

PIR:S22722 homeotic protein emx2 - human (fragment); N = 1, Score =
 763, P = 1.3e-75

25

TREMBL:0LA132403_1 gene: "emx2"; product: "Emx2 protein";
 Oryzias
 latipes mRNA for Emx2 protein, partial; N = 2, Score = 513, P =
 4.5e-72

30

>PIR:S22722 homeotic protein emx2 - human (fragment)
 Length = 158

HSPs:

35

Score = 763 (114.5 bits), Expect = 1.3e-75, P = 1.3e-75
 Identities = 144/144 (100%), Positives = 144/144 (100%)

Query: 109

40 FASQQRDPSTFYFWLIHRYRYLGHRFQGNDTSPEFLLHNALARKPKRIRTAFFSPSQLLR 168

FASQQRDPSTFYFWLIHRYRYLGHRFQGNDTSPEFLLHNALARKPKRIRTAFFSPSQLLR

Sbjct: 15

FASQQRDPSTFYFWLIHRYRYLGHRFQGNDTSPEFLLHNALARKPKRIRTAFFSPSQLLR 74

45

Query: 169

LEHAFEKNHYVVGAERKQLAHSLSLTETQVKVWFQNRRTKFKRQKLEEEGSDSQQKKKG 228

LEHAFEKNHYVVGAERKQLAHSLSLTETQVKVWFQNRRTKFKRQKLEEEGSDSQQKKKG

50 Sbjct: 75

LEHAFEKNHYVVGAERKQLAHSLSLTETQVKVWFQNRRTKFKRQKLEEEGSDSQQKKKG 134

Query: 229 HHINRWRIATKQASPEEIDVTSDD 252

HHINRWRIATKQASPEEIDVTSDD

55

Sbjct: 135 HHINRWRIATKQASPEEIDVTSDD 158

Pedant information for DKFZphamy2_14m1b, frame 1

Report for DKFZphamy2_14m1b.1

5 [LENGTH] 362
 [MW] 40749.28
 [pI] 10.51
 [HOMOL] PIR:I51737 homeotic protein emx2 - zebra fish 1e-113
 10 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YML027w] 5e-05
 [FUNCAT] 04.99 other transcription activities [S. cerevisiae, YML027w] 5e-05
 [FUNCAT] 03.07 pheromone response, mating-type determination, sex-specific proteins [S. cerevisiae, YCR097w] 5e-04
 15 [FUNCAT] 04.05.01.04 transcriptional control [S. cerevisiae, YDL106c] 7e-04
 [FUNCAT] 01.04.04 regulation of phosphate utilization [S. cerevisiae, YDL106c] 7e-04
 20 [FUNCAT] 01.03.13 regulation of nucleotide metabolism [S. cerevisiae, YDL106c] 7e-04
 [BLOCKS] PRO0049D
 [BLOCKS] PRO0909H
 25 [BLOCKS] PRO0487F
 [BLOCKS] PRO0796G
 [BLOCKS] BL00035C
 [BLOCKS] BL00027 'Homeobox' domain proteins
 [BLOCKS] PRO0026A
 30 [BLOCKS] BL00032C
 [BLOCKS] BL00032B 'Homeobox' antennapedia-type protein d1au7b1 1.4.1.1.6 Pit-1 POU homeodomain Pit-1 Pit-1 [Rat (Rattus 5e-16
 [SCOP] dlyrna_ 1.4.1.1.2 mating type protein A1
 35 Homeodomain mat alpha 2e-15
 [SCOP] dlenh_ 1.4.1.1.1 engrailed Homeodomain [(Drosophila melanogaster 2e-13
 [PIRKW] nucleus 1e-67
 [PIRKW] heart 3e-10
 40 [PIRKW] DNA binding 1e-67
 [PIRKW] leukemia 3e-15
 [PIRKW] alternative splicing 1e-10
 [PIRKW] proto-oncogene 3e-15
 [PIRKW] transcription factor 1e-11
 45 [PIRKW] embryo 9e-12
 [PIRKW] transcription regulation 1e-67
 [PIRKW] homeobox 1e-67
 [SUPFAM] homeobox homology 1e-67
 [SUPFAM] homeotic protein Hox A5 7e-10
 50 [SUPFAM] homeotic protein Hox B3 3e-10
 [SUPFAM] homeotic protein Hox B2 3e-11
 [SUPFAM] homeotic protein Hox B1 7e-11
 [SUPFAM] unassigned homeobox proteins 1e-67
 [SUPFAM] homeotic protein goosecoid 4e-10
 55 [SUPFAM] homeotic protein Hox D4 9e-12
 [PROSITE] HOMEBOX_1 1
 [PFAM] Homeobox domain
 [KW] Irregular

[KW]
[KW]

3D

LOW_COMPLEXITY 25.69 %

5 SEQ
EKKRKKKKKNYPNRLQILLEGFSPLSSGWARLVQDSRDPRLCLSLSLPPPFLFPFLSFL
SEG
.....
lfj1A

10

SEQ
SSFPSPHPHPKQTSPQFSSVLAAGSGRRRQRAAVARSWEPRAPAPRRSMFQAPAKRCF
SEG
15
lfj1A
.....

20 SEQ
TIESLVAKDSPLPASRSEDPIRPAALSYANSSPINPFLNGFHSAAAAAAGRGVYSNPDLV
SEG
.....
lfj1A
.....

25 SEQ
FAEAVSHPPNPAVPVHPVPPHALAAHPLPSSHSPHPLFASQQRDPSTFYPLIHRYRYL
SEG
.....
30 lfj1A
.....

35 SEQ
GHRFQGNDTSPESFLLHNALARKPKRIRTAFFSPSQLLRLEHAFEKNHYVVGAEKQLAHS
SEG
.....
lfj1A
.....CCCCCCCCCHHHHHHHHHHHHTTTTCHHHHHHHHHH

40 SEQ
LSLTETQVKVWFQNRRTKFKRQKLEEEGSDSQQKKKGTHHINRWRIATKQASPEEIDVTS
SEG
.....
lfj1A

45 HCCCHHHHHHHHHHHHHHHHHH.....

50 SEQ DD
SEG ..
lfj1A ..

Prositate for DKFZphamy2_14m1b.1

PS00027

297->321

HOME0BOX_1

PD0C00027

Pfam for DKFZphamy2_14m1b.1

HMM_NAME Homeobox domain

HMM

5 *RRRpRTtFTreQLdELEREFHfNrYPTRqRREELAQmLNLTERQVKIWF
+R RT+F+ +QL++LE +F+ N+Y+ ++R

+LA++L+LTE+QVK+WF

Query 264

10 PKRIRTAfSPSQLLRLEHAFEKNHYVVGAERKQLAHSLSLTETQVKVWF 312

HMM

QNRMMKWKRMM*

QNRK+K KR+

Query 313 QNRRTKFKRQK 323

15

DKFZphamy2_16e14

5 group: amygdala derived

DKFZphamy2_16e14.p3 encodes a novel 328 amino acid protein,
similar to carbonic anhydrase-related proteins.

10 A similar cDNA encoding a protein of the same length was
identified in sheep. This protein shows a strong signal sequence,
which indicates that it is a secreted protein. The new protein
belongs to a protein family, which was designated carbonic
15 anhydrase-related protein XI (CA-RP XI), encoded by CA11 (human)
and Car11 (mouse, rat). Despite potentially inactivating changes
in the active-site residues, CA-RP XI is evolving very slowly in
mammals, a property indicative of an important function, which
has also been observed in the two other "acatalytic" CA isoforms,
CA-RP VIII and CA-RP X.
20 No informative BLAST results; No predictive prosite, pfam or SCOP
motife.

The new protein can find application in studying the expression
profile of amygdala-specific genes.

25

similarity to carbonic anhydrase-related protein (Homo sapiens)

ESTs ending at appr. 1800 have polyA-signal

30

Sequenced by EMBL

Locus: /map="17q24; 5.13cR from GATA41C05"

35 Insert length: 2267 bp

Poly A stretch at pos. 2252, polyadenylation signal at pos. 2231

```

40  1 GGATGGAAAT AGTCTGGGAG GTGCTTTTTC TTCTTCAAGC CAATTTTCATC
    51 GTCTGCATAT CAGCTCAACA GAATTCACCA AAAATCCATG AAGGCTGGTG
   101 GGCATACAAG GAGGTGGTCC AGGGAAGCTT TGTTCAGTT CCTTCTTTCT
   151 GGGGATTGGT GAACTCAGCT TGGAACTCTT GCTCTGTGGG GAAACGGCAG
   201 TCGCCAGTCA ACATAGAGAC CAGTCACATG ATCTTCGACC CCTTTCTGAC
   251 ACCTCTTCGC ATCAACACGG GGGGCAGGAA GGTCAGTGGG ACCATGTACA
   301 ACACTGGAAG ACACGTATCC CTTGCCTGG ACAAGGAGCA CTTGGTCAAC
   351 ATATCTGGAG GGCCCATGAC ATACAGCCAC CGGCTGGAGG AGATCCGACT
   401 ACACTTTGGG AGTGAGGACA GCCAAGGGTC GGAGCACCTC CTCAATGGAC
   451 AGGCCTTCTC TGGGGAGGTG CAGCTCATCC ACTATAACCA TGAGCTATAT
   501 ACGAATGTCA CAGAAGCTGC AAAGAGTCCA AATGGATTGG TGGTAGTTTC
   551 TATATTTATA AAAGTTTCTG ATTCATCAAA CCCATTTCTT AATCGAATGC
   601 TCAACAGAGA TACTATCACA AGAATAACAT ATAAAAATGA TGCATATTTA
   651 CTACAGGGGC TTAATATAGA GGAACATAT CCAGAGACCT CTAGTTTCAT
   701 CACTTATGAT GGGTCGATGA CTATCCCACC CTGCTATGAG ACAGCAAGTT
   751 GGATCATAAT GAACAAACCT GTCTATATA CCAGGATGCA GATGCATTCC
   801 TTGCGCCTGC TCAGCCAGAA CCAGCCATCT CAGATCTTTC TGAGCATGAG
   851 TGACAACTTC AGGCCTGTCC AGCCACTCAA CAACCGCTGC ATCCGCACCA
   901 ATATCAACTT CAGTTTACAG GGGGAAGGACT GTCCAAACAA CCGAGCCAG
   951 AAGCTTCAGT ATAGAGTAAA TGAATGGCTC CTCAAGTAGG GAACAAAGCC

```



```

1001 AAGAAGAATC CCACCTCAGT GAAATGCTAC AACTGTGAAT TGACGTAACC
1051 TAGAATGTCC CCCTTCTTGC TTCTCTCTCC TTCTTTCCCC CAAGCCTCAT
1101 TCATTCTTGG GATTGGCCCT TTCTTCATGA AAAGTGTCTG CAAAACCATG
1151 GCAGAGGAAT ACATCTCTCA CACATACTCA CAAACACACA CACAAGCACT
5 1201 TGCACATACA TACAAACACA TGCAAAACATA CCTACACACA CACACACTCT
1251 TACAACCTCC ATCATGGGAA GTCAAGTTTC AGAAACAAAA GTCTCATTCA
1301 TAAGAGGTCT TAGAAGAAAA TAACCAGTTA ACCTGATTTT AATTTTGATA
1351 CCGTTTTTCTT GAACTAATAA ATCTACCCAA TGAGACTTTT CAGCCTTTGT
1401 ACATACAAAA TTCTTCCAAA AGAGAGAGGA GAAAATACAG CTCTGATGGC
10 1451 ATCAAACGGA CTTTGCATCA AGTAATTTCA GATAGTGTCC TAGGATCCTT
1501 TGAGGGGTGCT GGTAGCAGGT GAGCAGGACA AAGTTGACCA AGGACACTTA
1551 TTTCTAGATT ATGATTCTTC TGTTTACTCA ACAATTTACA AAGAAAAAAA
1601 GGACAGACAT TGAAGAGCTA CACATTGTAT ATATATCACC ACAGACTATA
1651 AGGAAATGGA ATTATTTCCC TCTTTGTCAC ATATCTGTAG TAGGATTTGC
15 1701 CAAGATCAGA AATGATCCAT TTGCTGTTTC TTGTTTTCCA AAGGTCATAC
1751 ATTGTGTTTG GTTATTGTTA CCAGCTCAAT AAATGTGTTT AACGAGTTAA
1801 TTTCATTTTT CTGGCTTTTG TCTGTTCTCC TTCCTTACAG GCTAAGCCCT
1851 GGCTCCATGC AACTGCATTC TTTGATTTC CTTGTTCCCT CATCTACATG
1901 TTTTGTTCAT TTGCAGCCAG TTTTACTGA GTTTGTGGCA ATCAGGAATG
20 1951 CATTTGCTAA GCAAGTATGA CTTTAATTCC ACTCCATGGC TCAATCATTC
2001 ACATGAGGTG AGCTTCAGCC TGAGATAGCA GGCAGACAGC TTCTTGCGTT
2051 TCAAAACTGC CATGCCCCC TGTGATGCTC CCGTGAAGGA ATGCACTTTG
2101 CCTTGTAAGT TCCTGGGAAA GGGGTATGTT TTCTCTCCAG GTGCAGCCAG
2151 ATCTACAAA GTACAAAACG AATGCCTTTC TTTTCTTGTT TATAATGGTC
25 2201 ACTCACTGTG TTTGGTTACT GTCAAGAAAT CAATAAATGT GTTTAACAAG
2251 TCAAAAAAAA AAAAAAA

```

BLAST alert EMBL/EMBLNEW

30

```

EMBL:AF064854 Homo sapiens map 17q24; 5.13cR from GATA41C05
repeat
region, complete sequence.; N = 2, Score = 8784, P = 0

```

35

```

EMBLNEW:AC005883 Homo sapiens chromosome 17 clone RP11-958E11 map
17,
WORKING DRAFT SEQUENCE, 2 ordered pieces.; N = 3, Score = 6260, P
= 0

```

40

Medline entries

9097349:

```

45 Lovejoy DA, Hewett-Emmett D, Porter CA, Cepoi D, Sheffield A,
Vale WW,

```

```

Tashian RE.; Evolutionarily conserved, "acatalytic" carbonic
anhydrase-related protein XI

```

```

50 contains a sequence motif present in the neuropeptide sauvagine:
the

```

```

human
CA-RP XI gene (CAL1) is embedded between the secretor gene
cluster and
the

```

```

55 DBP gene at 17q13.3. Genomics 1998 Dec 15;54(3):484-9

```

Peptide information for frame 3

5 ORF from 0 bp to 986 bp; peptide length: 329
 Category: similarity to known protein
 Classification: unclassified

```

10      1 MEIVWEVLFL LQANFIVCIS AQQNSPKIHE GWWAYKEVVQ GSFVPVPSFW
      51 GLVNSAWNLC SVGKRQSPVN IETSHMIFDP FLTPLRINTG GRKVS GMTMYN
      101 TGRHVSLRLD KEHLVNISGG PMTYSHRLEE IRLHFGSEDS QGSEHLLNGQ
      151 AFSGEVQLIH YNHELYTNVT EAAKSPNGLV VVSIFIKVSD SSNPFLNRML
      201 NRDTITRITY KNDAYLLQGL NIEELYPETS SFITYDGSMT IPPCYETASW
      251 IIMNKPVIIT RMQMHSRLRL SQNQPSQIFL SMSDNFRPVQ PLNNRCIRTN
15     301 INFSLQGKDC PNNRAQKLQY RVNEWLLK
  
```

Alert BLASTP hits for DKFZphamy2_16e14, frame 3

20 PIR:JED375 carbonic anhydrase-related protein - human; N = 1,
 Score =
 937, P = 4.6e-94

25 SWISSNEW:CAHB_SHEEP CARBONIC ANHYDRASE-RELATED PROTEIN 2
 PRECURSOR
 (CARP 2) (CA-RP II) (CA-XI); N = 1, Score = 935, P = 7.5e-94

30 >PIR:JED375 carbonic anhydrase-related protein - human
 Length = 328

HSPs:

35 Score = 937 (140.6 bits), Expect = 4.6e-94, P = 4.6e-94
 Identities = 169/287 (58%), Positives = 223/287 (77%)

Query: 30
 EGWWAYKEVVQGSFVPVPSFWGLVNSAWNLC SVGKRQSPVNIETSHMIFDPFLTPLRINT 89
 E WW+YK+ +QG+FVP P FWGLVN+AW+LC+VGKRQSPV++E

40 +++DPFL PLR++T
 Sbjct: 32
 EDWWSYKDNLQGNFVPGPPFWGLVNAAWSLC AVGKRQSPVDVEVKRVLYDPFLPPLRLST 91

Query: 90
 45 GGRKVS GMTMYNTGRHVSLRLDKEHLVNISGG PMTYSHRLEEIRLHFGSEDS QGSEHLLNG 149
 GG K+ GT+YNTGRHVS +VN+SGGP+ YSHRL E+RL FG+ D

GSEH +N
 Sbjct: 92
 GGEKLRGTLYNTGRHVSFLPAPRPVVNVSGGPLLYSHRLSELRLFLGARDGAGSEHQINH 151

50 Query: 150
 QAFSGEVQLIHYNHELYTNVTEAAKSPNGLVVVSIFIKVSDSSNPFLNRMLNRDTITRIT 209
 Q FS EVQLIH+N ELY N + A++ PNL ++S+F+ V+

+SNPFL+R+LNRDTITRI+
 55 Sbjct: 152
 QGFSAEVQLIHFNQELYGNFSAASRGPNGLAILSLFVNVASTSNPFLSRLNLRDTITRIS 211

Query: 210

YKNDAYLLQGLNIEELYPETSSFITYDGSMTIPPCYETASWIIMNKPVYITRMQMHSRL 269
 YKNDAY LQ L++E L+PE+ FITY GS++ PPC ET +WI++++ + IT
 +QMHSRL

5 Sbjct: 212

YKNDAYFLQDLSELELLFPESFGFITYQGSLSTPPCSETVTWILIDRALNITSLQMHSRL 271

Query: 270 LSQNPSPQIFLSMSDNFRPVQPLNNRCIRTNINFSLQGKDC--PNNR 314
 LSQN PSQIF S+S N RP+QPL +R +R N + + C PN R

10 Sbjct: 272 LSQNPSPQIFQSLSGNSRPLQPLAHRALRGNRDPRHPERRCRGPNYR 318

Pedant information for DKFZphamy2_1bcl4, frame 3

15

Report for DKFZphamy2_1bcl4.3

[LENGTH] 328
 [MW] 37563.19
 20 [pI] 8.22
 [HOMOL] PIR:JEO375 carbonic anhydrase-related protein -
 human 1e-101
 [BLOCKS] DM011098
 [BLOCKS] BL00162F
 25 [BLOCKS] BL00162E
 [BLOCKS] BL00162D
 [BLOCKS] BL00162C Eukaryotic-type carbonic anhydrases
 proteins
 [BLOCKS] BL00162A Eukaryotic-type carbonic anhydrases
 30 proteins
 [SCOP] dlznca_ 2.5b.1.1.3 Carbonic anhydrase [human
 (Homo sapiens 1e-103
 [SCOP] d2cba_ 2.5b.1.1.2 Carbonic anhydrase [human
 (Homo sapiens 9e-97
 35 [EC] 4.2.1.1 Carbonate dehydratase 1e-36
 [EC] 3.1.3.48 Protein-tyrosine-phosphatase 2e-20
 [PIRKW] blocked amino end 8e-29
 [PIRKW] carbon-oxygen lyase 1e-36
 [PIRKW] zinc 1e-36
 40 [PIRKW] polymorphism 2e-20
 [PIRKW] hydro-lyase 1e-36
 [PIRKW] transmembrane protein 3e-23
 [PIRKW] tyrosine-specific phosphatase 2e-20
 [PIRKW] brain 6e-16
 45 [PIRKW] acetylated amino end 1e-36
 [PIRKW] phosphatidylinositol linkage 2e-19
 [PIRKW] receptor 2e-20
 [PIRKW] liver 3e-29
 [PIRKW] phosphoprotein 2e-20
 50 [PIRKW] saliva 2e-21
 [PIRKW] glycoprotein 2e-22
 [PIRKW] mitochondrion 1e-32
 [PIRKW] monomer 3e-32
 [PIRKW] alternative splicing 6e-16
 55 [PIRKW] lipoprotein 2e-19
 [PIRKW] pyroglutamic acid 2e-21
 [PIRKW] metalloprotein 6e-35
 [PIRKW] muscle 4e-31

5 [PIRKW] membrane protein 2e-19
 [PIRKW] phosphoric monoester hydrolase 2e-20
 [PIRKW] homodimer 3e-23
 [SUPFAM] fibronectin type III repeat homology 2e-20
 [SUPFAM] carbonic anhydrase homology 1e-36
 [SUPFAM] protein-tyrosine-phosphatase, receptor type zeta
 1e-16
 [SUPFAM] carbonate dehydratase 1e-36
 [SUPFAM] protein-tyrosine-phosphatase, receptor type gamma
 10 2e-20
 [SUPFAM] protein-tyrosine-phosphatase homology 2e-20
 [SUPFAM] leukocyte common antigen cytosolic domain
 homology 2e-20
 [PFAM] Eukaryotic-type carbonic anhydrases
 15 [KW] All_Beta
 [KW] 3D
 [KW] SIGNAL_PEPTIDE 22

20 SEQ
 MEIVWEVLFLLLQANFIVCISAQQNSPKIHEGWWAYKEVVQGSFVPVPSFWGLVNSAWNLC
 luc-

25 SEQ
 SVGKRQSPVNIETSHMIFDPFLTPLRINTGGRKVS GMTMYNTGRHVSLRLDKEHLVNISGG
 luc- ..TTTCCCEETTTTTEETTTTCEEEETT-
 TTCEEEETTTTTEEEECTTTTTEEEEE

30 SEQ
 PMTYSHRLEEIRLHFGSEDSQGEHLLNGQAFSGEVQLIHYNHLYTNVTEAAKSPNGLV
 luc- TTCCCEEEEEEEEEETTTTTTCTTTEETTBCCCEEEEEEEEEEGG-
 GTTHHHHCTTTTEE

35 SEQ
 VVSIFIKVSDSSNPFLNRMLNRDTITRITYKNDAYLLQGLNIEELYPETSSFITDGSMT
 luc- EEEEEEEEC-CCCGGGHHHH--
 HHGGGCCCTTTEEEETTTTCGGGGCCCCCEEEEECCC

40 SEQ
 IPPCYETASWIIMNKPVIITRMQMSLRLLSQNQPSQIFLSMSDNFRPVQPLNNRCIRTN
 luc-
 TTTTCCCEEEEEEECCCEEECHHHHHHHHCCBCCTTTTCCCBTTTTCCCCCTTTTCCEEC

45 SEQ INFSLQGKDCPNNRAQKLQYRVNEULLK
 luc-

50 (No Prosite data available for DKFZphamy2_1b14.3)
 Pfam for DKFZphamy2_1b14.3

55 HMM_NAME Eukaryotic-type carbonic anhydrases
 HMM
 *WCYgeHWGPEHH.....WHkhYPIAW....GDRQSPINIQWkearyYDPS

W Y E +

W+++ + +

G RQSP+NI ++

+DP

Query

33

WAYKEVVQGSFVPVPSFWGLVNSAWNLC SVGKRQSPVNIETSHMIFDPF 81

5

HMM

LKPWrv.SYYpaWCrEWeIWNNGHSFQVeFDDSDMSVLsGGPLPgHPYR

L P+R+ ++ +++++ ++ N+G+ + +D

+SGGP++

++R

10

Query

82

LTPLRINTGGRKVS--TMYNTGRHVSLRLDK-
EHLVNISGGPMTY-SHR 127

HMM

LKQFHFHUGGASSNDWGSEHTVDGmkYPMEHLHVWNStKYnNYdEAQdq

L + ++H G S++ +GSEH ++G +++ E+ L+H+N +Y N+

15

EA++

Query

128

LEEIRLHFG--
SEDSQGEHLLNGQAFSGEVQLIHYNHELYTNVTEAAKS 175

20

HMM

PDGLAVIGVFMKVGNYqENPyLQKVv.-DALdnIKYKGKratMTNFDPsC

P+GL V+ +F+KV NP L++ + D + I YK +

+++++

Query

176

PNGLVVVSIFIKVS-
DSSNPFLNRMLNRDTITRITYKNDAYLLQGLNIEE 224

25

HMM

LLPpPnCRDYWTYPGSLTTPPCHECVTWIVCKEPIsISsEQMWKFRsLLF

L P+ + TY GS+T+PPC+E WI+ P+ I + QM +R

30

L

Query

225

LYPE--
TSSFITYDGSMTIPPCYETASWIIMNKPVYITRMQMSLRLLSQ 272

HMM

NHEGEeeVpMVDNWRRPQPLKhRvVRASF*

N +M DN+RP QPL++R +R +

35

Query

273

NQPSQIFLSMSDNFRPVQPLNNRCIRTNI 301

DKFZphamy2_1c12

5 group: nucleic acid management

DKFZphamy2_1c12 encodes a novel 422 amino acid protein with partial identity to I-kappa-B-related protein and to BRCA1.

10 I-kappa-B-related protein interacts with transcription factors and BRCA1 has a function in DNA damage response. I-kappa-B-alpha mutations contribute to constitutive NF-kappaB activity in cultured and primary HRS (Hodgkin/Reed-Sternberg) cells and are therefore involved in the pathogenesis of Hodgkin's disease (HD) patients.

The new protein can find application in modulating DNA repair and mutagenesis and also in expression profiling in HD related syndroms.

similarity to I-kappa-B-related protein

Sequenced by MediGenomix

Locus: unknown

Insert length: 1645 bp

Poly A stretch at pos. 1626, polyadenylation signal at pos. 1605

```

1  GGATTTTCCT  TGGTCTTAAG  ATGGGTAGAA  ATGTGATGCG  ACACATGTCT
51  GATGACTTAG  GAAGTTATGT  TTCTCTTTTC  TGTGATGACT  TTTCTTCACA
101  GGAATTAGAG  ATTTTCATTT  GCTCCTTTTC  CTCCTCCTGG  CTTCAAATGT
35  151  TTGTTGCAGA  GGCAGTCTTT  AAAAAGTTGT  GTCTACAGAG  CTCTGGCAGT
201  GTTTCTTCTG  AGCCACTCTC  TCTTCAGAAA  ATGGTATATT  CCTATTTACC
251  AGCCTTGGGG  AAAACTGGTG  TGCTTGGGTC  TGGAAAGATT  CAGGTGTCAA
301  AGAAAATAGG  ACAGCGGCCT  TGTTTTGACT  CTCAGAGAAC  CTTACTAATG
351  CTGAATGGTA  CTAAACAAAA  ACAAGTCGAA  GGGCTGCCAG  AGTTACTAGA
40  401  CCTGAACCTT  GCTAAATGTT  CCTCATCATT  AAAAAAATTG  AAAAAGAAGT
451  CAGAAGGAGA  ATTGTCATGT  TCCAAGGAGA  ATTGCCCTC  TGTAGTTAAA
501  AAGATGAATT  TTCACAAGAC  TAATCTAAAA  GGAGAAACAG  CCCTGCATAG
551  AGCTTGCATA  AATAACCAAG  TGGAGAAATT  GATTCTTCTT  CTCTCTTTGC
601  CAGGAATAGA  CATCAATGTT  AAAGACAATG  CTGGCTGGAC  GCCTTTGCAT
45  651  GAAGCCTGTA  ACTATGGCAA  CACAGTGTGT  GTCCAGGAAA  TTTTGCAACG
701  TTGTCCAGAG  GTAGATCTGC  TCACTCAAGT  GGACGGGGTG  ACTCCTTTGC
751  ATGATGCACT  GTCAAACGGA  CATGTAGAAA  TTGGCAAGCT  GCTACTACAG
801  CATGGGGGCC  CAGTGCTTTT  ACAACAGAGG  AATGCTAAGG  GAGAATTGCC
851  CTTGGATTAT  GTGGTTTCAC  CTCAAATCAA  AGAAGAACTG  TTTGCTATTA
50  901  CAAAATAGA  AGATACAGTG  GAGAACTTTC  ATGCACAAGC  AGAGAAACAT
951  TTTCATTACC  AGCAACTTGA  ATTTGGCTCC  TTTTACTTAA  GTAGGATGTT
1001  GCTAAATTTT  TGTTCAATTT  TTGATTTATC  TTCAGAGTTC  ATTTTAGCTT
1051  CCAAAGGGTT  AACTCATCTA  AATGAACTGC  TTATGGCTTG  TAAAAGTCAT
1101  AAAGAAACCA  CCAGTGTTCA  TACTGACTGG  TTA CTGGATC  TTTATGCTGG
55  1151  AAATATAAG  ACATTGCAGA  AACTCCCACA  CATTCTTAAG  GAAC TGCCTG
1201  AGAATTTGAA  AGTGTGTCCT  GGGGTACACA  CTGAGGCCTT  GATGATAACA
1251  TTGGAAATGA  TGTGTCGGTC  AGTCATGGAG  TTTTCATGAT  GATGCTAGAA
1301  AGTATGGATT  GACTTTCTAA  ATCTGTTTCA  TTTGCATTGG  TACTTACTGT

```

1351 GGACTTCATA GCTTACTGAC AGATAGTAAT TTGATTTATT TATTGACAGA
 1401 CTTTGCAGCC TTGCTAAATT TTAAGAGCAT TTTTAAAAAA ACTTCTACAA
 1451 AACTCTAGTA TGGGCTTCTG ACTTTTTCCTA GGGTGCTAGAA TTTGACTCAA
 1501 AAGTAAAAAT AATTTTGTTC TAGTATATTC TACTTTCATT AATGTTTTTT
 5 1551 TGTTCCTGAAA GTGATATTAT ATTGTACATG TAAAATTAAT TTTAAATATT
 1601 TTTCAAATAA AAATGTAATG TCCTGTAAAA AAAAAAAAAA AAAAA

BLAST Results

10

No BLAST result

15

Medline entries

No Medline entry

20

Peptide information for frame 3

25 ORF from 21 bp to 1286 bp; peptide length: 422
 Category: similarity to known protein
 Classification: Cell signaling/communication

30 1 MGRNVMRHMS DDLGSYVSLG CDDFSSQELE IFICSFSSSW LQMFVAEAVF
 51 KKLCLQSSGS VSSEPLSLQK MVYSYLPALG KTGVLGSGKI QVSKKIGQRP
 101 CFDSQRTLLM LNGTKQKQVE GLPELLDLNL AKCSSSLKKL KKKSEGELSC
 151 SKENCPSVVK KMNFKHTNLK GETALHRACI NNQVEKLILL LSLPGIDINV
 201 KDNAGWTPLH EACNYGNTVC VQELQRCPE VDLLTQVDGV TPLHDALNSG
 251 HVEIGKLLLQ HGGPVLLQQR NAKGELPLDY VVSPQIKEEL FAITKIEDTV
 35 301 ENFHAQAEKH FHYQQLFEGS FLLSRMLLNF CSIFDLSEF ILASKGLTHL
 351 NELLMAKSH KETTSVHTDW LLDLYAGNIK TLQKLPHILK ELPENLKVCP
 401 GVHTEALMIT LEMMCRSVME FS

40

BLASTP hits

No BLASTP hits available

45

Alert BLASTP hits for DKFZphamy2_1c12, frame 3

PIR:A56429 I-kappa-B-related protein - human; N = 1, Score = 242,
 P =
 4.6e-18

50

TREMBLNEW:AF038042_1 gene: "BARD1"; product: "BRCA1-associated
 RING
 domain protein"; Homo sapiens BRCA1-associated RING domain
 protein

55

(BARD1) gene, exons 10, 11 and complete cds.; N = 1, Score = 236,
 P =
 6.7e-17

>PIR:A5b429 I-kappa-B-related protein - human
Length = 481

5 HSPs:

Score = 242 (36.3 bits), Expect = 4.6e-18, P = 4.6e-18
Identities = 52/118 (44%), Positives = 71/118 (60%)

10 Query: 156
PSVVKKMNFHKTNLKGETALHRACINNQVEKLILLLSLPGLDINVKDNAGWTPLEACNY 215
P K +++ N GET LHRACI Q+ ++ L+ G +N +D
WTPLEACNY
Sbjct: 354 PGAAGGSKWNRRNDMGETLLHRACIEGQLRRVQDLVR-
15 QGHPLNPRDYCGWTPLEACNY 412

Query: 216 GNTVCVQIEILQRCPEVDLL--
TQVDGVTPLHDALSNHVEIGKLLLQHGPPVLLQQRNA 272
G+ V+ +L VD +G+TPLHDAL+ GH E+ +LLL+ G V
20 L+ R A
Sbjct: 413
GHLEIVRFLLDHGAAYDDPGGQGCEGITPLHDALNCGHFEVAELLERGASVTLRTRKA 471

25 Pedant information for DKFZphamy2_1c12, frame 3

Report for DKFZphamy2_1c12.3

30 [LENGTH] 422
[MW] 47071.18
[pI] 6.57
[HOMOL] PIR:A5b429 I-kappa-B-related protein - human 3e-19
35 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YIL112w]
3e-11
[FUNCAT] 06.13.01 cytoplasmic degradation [S. cerevisiae,
YGR232w] 4e-06
40 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YIR033w]
2e-04
[FUNCAT] 04.05.01.07 chromatin modification [S. cerevisiae,
YIR033w] 2e-04
[SCOP] dlawcb_ 1.91.3.1.2 GA binding protein (GABP) alpha
45 GA binding 6e-24
[EC] 3.1.3.53 Myosin-light-chain-phosphatase 9e-06
[PIRKW] phosphotransferase 3e-07
[PIRKW] tandem repeat 9e-06
[PIRKW] transmembrane protein 7e-10
50 [PIRKW] serine/threonine-specific protein kinase 3e-07
[PIRKW] phosphoprotein 3e-10
[PIRKW] integrin binding 3e-07
[PIRKW] alternative splicing 3e-11
[PIRKW] peripheral membrane protein 2e-09
55 [PIRKW] transcription regulation 3e-06
[PIRKW] phosphoric monoester hydrolase 9e-06
[PIRKW] cytoskeleton 4e-10
[PIRKW] smooth muscle 9e-06

[SUPFAM] ankyrin 3e-11
 [SUPFAM] ankyrin repeat homology 3e-11
 [SUPFAM] unassigned ankyrin repeat proteins 7e-10
 [PFAM] Ank repeat
 5 [KW] Irregular
 [KW] 3D
 [KW] LOW_COMPLEXITY 8.53 %

10 SEQ MGRNVMRHMSDDLGSYVSLSCDDFSSQELEIFICSFSSSWLQMFVAEAVFKKLCLOSSGS
 SEGxxxxxx
 lawcB

15 SEQ VSSEPLSLQKMVYSYLPALGKTGVLGSGKIQVSKKIGQPCFDSQRTLLMLNGTKQKQVE
 SEG xxxxxxxx.....
 lawcB

20 SEQ GLPELLDLNLAKCSSSLKKLKKKSEGELSCSKENCPSVVKKMNFHKTNLKGETALHRACI
 SEGxxxxxxxxxxxxxxxxxxxxxxxx.....
 lawcB

25 SEQ NNQVEKLILLLSLPGIDINVKDNAGWTPLEACNYGNTVCVQEILQRCPEVDLLTQVDGV
 SEG
 lawcBTTTTTCHHHHHHHHCCHHHHHHHHHCCCTTTTCTTTTC-

30 SEQ TPLHDALSNGHVEIGKLLLQHGPPVLLQQRNAKGELPLDYVVSPPQIKEELFAITKIEDTV
 SEG
 lawcB CHHHHHHHHTTHHHHHHHHHCCCTT.....

35 SEQ -ENFHAQAEKHFHYQQLFSGSFLLSRMLLNFCISIFDLSSSEFILASKGLTHLNELLMACKSH
 SEG
 lawcB

40 SEQ KETTSVHTDWLLDLYAGNIKTLQKLPHILKELPENLKVCPGVHTEALMITLEMMCRSVME
 SEG
 lawcB

45 SEQ FS
 SEG ..
 lawcB ..

50 (No Prosite data available for DKFZphamy2_1c12.3)

Pfam for DKFZphamy2_1c12.3

55

HMM_NAME Ank repeat

HMM *GyTPLHIAARyNNvEMVrILLQH.GADIN*

Query 171 G+T+LH A+++N+VE LLL+ G DIN
GETALHRACINNQVEKLILLLSLPGIDIN 199

34.48 (bits) f: 205 t: 232 Target: dkfzphamy2_1c12.3
similarity to I-kappa-B-related protein

Alignment to HMM consensus:

Query *GYTPLHIAARyNNvEMVr1LLQHGADIN*
G+TPLH A+ Y+N+ +V+ LQ+ + ++
dkfzphamy2 205 GWTPLHEACNYGNTVCVQEIILQRCPEVD 232

Query f: 239 t: 266 Target: dkfzphamy2_1c12.3
similarity to I-kappa-B-related protein

Alignment to HMM consensus:

HMM *GYTPLHIAARyNNvEMVr1LLQHGADIN*
G TPLH A +++VE+ +LLLQHG +
Query 239 GVTPLHDALSNGHVEIGKLLLQHGGPVL 266

DKFZphamy2_1i1

group: nucleic acid management

DKFZphamy2_1i1 encodes a novel 629 amino acid protein with
similarity to the murine hemin-sensitive initiation factor 2.

The hemin-sensitive initiation factor 2 is expressed
predominantly in liver, spleen, colon and uterus and contains 2
protein kinase motifs. The mouse homologue inhibits protein
synthesis in stress conditions by phosphorylation of eif-2-alpha.
Four different eIF2alpha kinases have been identified in
mammalian cells, the heme-regulated inhibitor (HRI), the
interferon-inducible RNA-dependent kinase (PKR), the endoplasmic
reticulum-resident kinase (PERK) and MGCN2. The new protein
represents a new member of this family

The new protein can find application in modulating/blocking of
translation.

similarity to hemin-sensitive initiation factor 2 (Mus musculus),
complete cds.alpha kinase

complete cds.

probably complete in genomic clone DJ0042M02

Sequenced by MediGenomix

Locus: /map="37.2 cR from top of Chr7 linkage group"

Insert length: 2863 bp

Poly A stretch at pos. 2844, polyadenylation signal at pos. 2824

1 GCAGTGCTGG GCTGGCCGGC GGGCTGGGCT GCGGCCCGCG CGCGGCCGGC
51 GATGCAGGGG GGCAACTCCG GGGTCCGCAA GCGCGAAGAG GAGGGCGACG
101 GGGCTGGGGC TGTGGCTGCG CCGCCG6CCA TCGACTTTCC CGCCGAGGGC
151 CCGGACCCCG AATATGACGA ATCTGATGTT CCAGCAGAAA TCCAGGTGTT

201 AAAAGAACCC CTACAACAGC CAACCTTCCC TTTTGCAGTT GCAAACCAAC
251 TCTTGCTGGT TTCTTTGCTG GAGCACTTGA GCCACGTGCA TGAACCAAAC
301 CCACTTCGTT CAAGACAGGT GTTTAAGCTA CTTTGCCAGA CGTTTATCAA
351 AATGGGGCTG CTGTCTTCTT TCACTTGTAG TGACGAGTTT AGCTCATTGA
5 401 GACTACATCA CAACAGAGCT ATTACTCACT TAATGAGGTC TGCTAAAGAG
451 AGAGTTCGTC AGGATCCTTG TGAGGATATT TCTCGTATCC AGAAAATCAG
501 ATCAAGGGAA GTAGCCTTGG AAGCACAAAC TTCACGTTAC TTAAATGAAT
551 TTGAAGAACT TGCCATCTTA GGAAGAGGTG GATACGGAAG AGTATACAAG
601 GTCAGGAATA AATTAGATGG TCAGTATTAT GCAATAAAAA AAATCCTGAT
10 651 TAAGGGTGCA ACTAAACAG TTTGCATGAA GGTCTACGG GAAGTGAAGG
701 TGCTGGCAGG TCTTCAGCAC CCCAATATTG TTGGCTATCA CACCGCGTGG
751 ATAGAACATG TTCATGTGAT TCAGCCACGA GACAGAGCTG CCATTGAGTT
801 GCCATCTCTG GAAGTGCTCT CCGACCAGGA AGAGGACAGA GAGCAATGTG
851 GTGTTAAAAA TGATGAAAGT AGCAGCTCAT CCATTATCTT TGCTGAGCCC
15 901 ACCCCAGAAA AAGAAAAACG CTTTGGAGAA TCTGACACTG AAAATCAGAA
951 TAACAAGTCG GTGAAGTACA CCACCAATTT AGTCATAAGA GAATCTGGTG
1001 AACTTGAGTC GACCCTGGAG CTCCAGGAAA ATGGCTTGGC TGGTTTGTCT
1051 GCCAGTTCAA TTGTGGAACA GCAGCTGCCA CTCAGGCGTA ATTCCCACCT
1101 AGAGGAGAGT TTCACATCCA CCGAAGAATC TTCCGAAGAA AATGTCAACT
20 1151 TTTTGGGTCA GACAGAGGCA CAGTACCACC TGATGCTGCA CATCCAGATG
1201 CAGCTGTGTG AGCTCTCGCT GTGGGATTGG ATAGTCGAGA GAAACAAGCG
1251 GGGCCGGGAG TATGTGGACG AGTCTGCCTG TCCTTATGTT TCCGCCAATG
1301 TTGCAACAAA AATTTTTCOA GAATTGGTAG AAGGTGTGTT TTACATACAT
1351 AACATGGGAA TTGTGCACCG AGATCTGAAG CCAAGAAATA TTTTCTTCA
25 1401 TGGCCCTGAT CAGCAAGTAA AAATAGGAGA CTTTGGTCTG GCCTGCACAG
1451 ACATCCTACA GAAGAACACA GACTGGACCA ACAGAAACGG GAAGAGACA
1501 CCAACACATA CGTCCAGAGT GGGTACTTGT CTGTACGCTT CACCCGAACA
1551 GTTGGGAAGGA TCTGAGTATG ATGCCAAGTC AGATATGTAC AGCTTGGGTG
1601 TGGTCCTGCT AGAGCTCTTT CAGCCGTTTG GAACAGAAAT GGGCCAGCA
30 1651 GAAGTTCTAA CAGGTTTAAAG AACTGGTCAG TTGCCGGAAT CCCTCCGTAA
1701 AAGGTGTCCA GTGCAAGCCA AGTATATCCA GCACTTAACG AGAAGGAACT
1751 CATCGCAGAG ACCATCTGCC ATTCAGCTGC TGCAGAGTGA ACTTTTCCAA
1801 AATTCTGGAA ATGTAAACCT CACCCTACAG ATGAAGATAA TAGAGCAAGA
1851 AAAAGAAATT GCAGAACTAA AGAAGCAGCT AAACCTCCTT TCTCAAGACA
35 1901 AAGGGGTGAG GGTGACGGA AAGGATGGGG GCGTGGGATG AAAGTGGACT
1951 TAACTTTTAA GGTAGTTAAG TGGAATGTAA ATTTTAAATC TTTATTAGGG
2001 TATAGTTGGT ACAATGCTTC GTTGATTATA GTAAGCCTTT ACAAGACTTG
2051 TTAAAGATGT CAGAGTGCCC CAAGCTGCCG TTCCTTCCCT TCCTGCCCCA
40 2101 CAAGCTCCTT TTCCTGAATT TCCTACCTAA ATATTAACCA TATGCCTAGT
2151 CTCTGAAACT AAAAACTTGG ACCTCATCCT CAATTATTTT CTCCTTTCAA
2201 CTCTGTTGAC CCTCTGTCTG GTCTTCCTCT AGAAGGTTCT ACCGCAGAAA
2251 TTGATGTGTG CTCCCTGCCC TCGTCACTGC CCAAGCCCGG GCCTGCACAT
2301 ACTCACTGGA CTGTTCCAGT TTTGACAGCT GCCAGTCTTC CTGCCCTTTT
2351 CACACTGCAG CTGAAGTTCA TTACCTGAAG GACGCCTCAT CATTTTATTTC
45 2401 CTTGGCTCCA AACCTTCTGC TGCCTTAAG ATAAAAGCTC AACTTCTTAA
2451 CAGTGTACAG TGTGCAACTT CCAACCTTTT TATCTGTTCT CTCCACCTTC
2501 AGTTTAGCGT CATTCAAAA CCACACCCTT GCAAAGCTTT GTACTCCGCA
2551 CCCCAGATGA TCTCCAGGCA GCTCAGATCT CTTTCCTGCC TTTGCCCTGC
2601 ACTGTTCCCC GGTACTTCCT CTTTATTGT AGCACTCAGC TCCCCAGCCA
50 2651 ATCTGTACAT CCCTCAGAGG CAGCGATCTG ATGAATTGGT TTTTGAATCC
2701 CAGAAAGGGT CTGCCATGGA GTTGGCAGTC ATCACGGTAG ATGGCGTATG
2751 ATTTTGCTGA ATTTTAAATA AAATGAAAC CATAAATTAC ATGATGCTTT
2801 TATTGACACT TGACAACTGG CCTAAATAAA AAGACTCTGA CTCCAAAAAA
2851 AAAAAAAAAA AAA

BLAST Results

Entry AF028808 from database EMBL:
Mus musculus hemin-sensitive initiation factor 2 alpha kinase
mRNA,

5 complete cds.

Score = 6688, P = 2.7e-296, identities = 1922/2534

Entry AC005995 from database EMBL:

10 Homo sapiens clone DJ0042M02, WORKING DRAFT SEQUENCE, 13
unordered
pieces.

Score = 5116, P = 0.0e+00, identities = 1090/1148

15

Medline entries

99042009:

20 Berlanga J.J., Herrero S., de Haro C.; Characterization of the
hemin-sensitive eukaryotic initiation factor 2alpha kinase from
mouse
nonerythroid cells; J. Biol. Chem. 273(48):32340-32346(1998).

25

Peptide information for frame 1

30

ORF from 52 bp to 1938 bp; peptide length: 629

Category: similarity to known protein

Classification: Protein management

Prosite motifs: PROTEIN_KINASE_ATP (173-196)

35 PROTEIN_KINASE_ATP (173-197)

PROTEIN_KINASE_ST (437-449)

40

1	MQGGNSGVRK	REEEGDGAGA	VAAPPAIDFP	AEGPDPEYDE	SDVPAEIQVL
51	KEPLQPTFP	FAVANQLLV	SLLEHLSHVH	EPNPLRSRQV	FKLLCQTFIK
101	MGLLSSTCS	DEFSSLRLHH	NRAITHLMRS	AKERVQRDPC	EDISRIQKIR
151	SREVALEAQT	SRYLNEFEEL	AILGKGGYGR	VYKVRNKLDG	QYYAIKKILI
201	KGATKTVCMK	VLREVKVLAG	LQHPNIVGYH	TAWIEHVHVI	QPRDRAAIEL
251	PSLEVLSDAQE	EDREQCGVKN	DESSSSSIIF	AEPTPEKEKR	FGESDTENQN
301	NKSVKYTTNL	VIRESGELES	TLELQENGLA	GLSASSIVEQ	QLPLRRNSHL
351	EESFTSTEEES	SEENVNFLGQ	TEAQYHMLH	IQMQLCELSL	WDWIVERNKR
401	GREYVDESAC	PYVMANVATK	IFQELVEGVF	YIHNMGIVHR	DLKPRNIFLH
451	GPDQQVKIGD	FGLACTDILQ	KNTDWTNRNG	KRTPHTSRV	GTCLYASPEQ
501	LEGSEYDAKS	DMYSLGVVLL	ELFQPFGTET	ERAEVLTGLR	TGQLPESLRK
551	RCPVQAKYIQ	HLTRRNSSQR	PSAIQLLQSE	LFQNSGNVNL	TLQMKIIEQE
601	KEIAELKKQL	NLLSQDKGVR	DDGKDGGVG		

55

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_lil, frame 1

No Alert BLASTP hits found

5 Pedant information for DKFZphamy2_lil, frame 1

Report for DKFZphamy2_lil.1

10

[LENGTH] 646
 [MW] 72738.78
 [pI] 5.80
 [HOMOL] SWISSNEW:HRI_MOUSE HEME-REGULATED EUKARYOTIC
 15 INITIATION FACTOR EIF-2-ALPHA KINASE (EC 2.7.1.-) (HEME-REGULATED
 INHIBITOR) (HRI) (HEME-CONTROLLED REPRESSOR) (HCR) (HEMIN-
 SENSITIVE INITIATION FACTOR-2 ALPHA KINASE). 0.0
 [FUNCAT] 05.07 translational control [S. cerevisiae, YDR283c] 2e-43
 20 [FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae, YDR283c] 2e-43
 [FUNCAT] 10.02.11 key kinases [S. cerevisiae, YOR231w] 8e-14
 [FUNCAT] 03.04 budding, cell polarity and filament formation [S. cerevisiae, YOR231w] 8e-14
 25 [FUNCAT] 03.01 cell growth [S. cerevisiae, YOR231w] 8e-14
 [FUNCAT] 11.01 stress response [S. cerevisiae, YOR231w] 8e-14
 [FUNCAT] 03.22 cell cycle control and mitosis [S. cerevisiae, YOR231w] 8e-14
 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YKL101w] 8e-12
 30 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YPL150w] 8e-12
 [FUNCAT] 03.13 meiosis [S. cerevisiae, YDR523c] 2e-11
 [FUNCAT] 03.10 sporulation and germination [S. cerevisiae, YDR523c] 2e-11
 35 [FUNCAT] 09.01 biogenesis of cell wall [S. cerevisiae, YPL140c] 4e-11
 [FUNCAT] 10.03.11 key kinases [S. cerevisiae, YCR073c] 9e-11
 [FUNCAT] 98 classification not yet clear-cut [S. cerevisiae, YHR082c] 1e-10
 40 [FUNCAT] 03.07 pheromone response, mating-type determination, sex-specific proteins [S. cerevisiae, YLR362w] 2e-10
 [FUNCAT] 10.05.11 key kinases [S. cerevisiae, YLR362w] 2e-10
 [FUNCAT] 10.04.11 key kinases [S. cerevisiae, YLR362w] 2e-10
 45 [FUNCAT] 10.99 other signal-transduction activities [S. cerevisiae, YDL101c] 3e-10
 [FUNCAT] 11.04 dna repair (direct repair, base excision repair and nucleotide excision repair) [S. cerevisiae, YDL101c] 3e-10
 [FUNCAT] 03.25 cytokinesis [S. cerevisiae, YDR507c] 3e-10
 50 [FUNCAT] 04.05.01.01 general transcription activities [S. cerevisiae, YDL108w] 1e-09
 [FUNCAT] 03.16 dna synthesis and replication [S. cerevisiae, YBR160w] 1e-09
 [FUNCAT] 01.05.04 regulation of carbohydrate utilization [S. cerevisiae, YLR113w] 4e-09
 55 [FUNCAT] 02.19 metabolism of energy reserves (glycogen, trehalose) [S. cerevisiae, YPL031c] 1e-08

- 1 [FUNCAT] 04.05.01.04 transcriptional control [S. cerevisiae, YPL031c] 1e-08
 [FUNCAT] 01.04.04 regulation of phosphate utilization [S. cerevisiae, YPL031c] 1e-08
 5 [FUNCAT] c energy conversion [M. genitalium, MG109] 2e-08
 [FUNCAT] 03.19 recombination and dna repair [S. cerevisiae, YOR351c] 1e-07
 [FUNCAT] 03.22.01 cell cycle check point proteins [S. cerevisiae, YPL153c] 1e-07
 10 [FUNCAT] 10.05.09 regulation of g-protein activity [S. cerevisiae, YBL016w] 7e-07
 [FUNCAT] 04.03.99 other trna-transcription activities [S. cerevisiae, YIL035c] 1e-06
 [FUNCAT] 08.13 vacuolar transport [S. cerevisiae, YGL180w] 1e-06
 15 [FUNCAT] 06.13.04 lysosomal and vacuolar degradation [S. cerevisiae, YGL180w] 1e-06
 [FUNCAT] 04.99 other transcription activities [S. cerevisiae, YER129w] 2e-06
 20 [FUNCAT] 30.02 organization of plasma membrane [S. cerevisiae, YDR122w] 2e-06
 [FUNCAT] 30.07 organization of endoplasmatic reticulum [S. cerevisiae, YHR079c] 3e-06
 [FUNCAT] 01.06.10 regulation of lipid, fatty-acid and sterol biosynthesis [S. cerevisiae, YHR079c] 3e-06
 25 [FUNCAT] 08.99 other intracellular-transport activities [S. cerevisiae, YKL198c] 1e-05
 [FUNCAT] 10.04.99 other nutritional-response activities [S. cerevisiae, YKL198c] 1e-05
 30 [FUNCAT] 09.04 biogenesis of cytoskeleton [S. cerevisiae, YNL020c] 9e-05
 [FUNCAT] 06.07 protein modification (glycosylation, acylation, myristylation, palmitylation, farnesylation and processing) [S. cerevisiae, YFLO33c] 4e-04
 35 [FUNCAT] 01.02.04 regulation of nitrogen and sulphur utilization [S. cerevisiae, YNL183c] 7e-04
 [BLOCKS] BLO0107A Protein kinases ATP-binding region proteins [SCOP] d1ir3a_ 5.1.1.2.6 insulin receptor Complex (transferase/substrate) 1e-22
 40 [SCOP] d1fgkb_ 5.1.1.2.5 Fibroblast growth factor receptor 1 [Human (Homo)] 9e-27
 [SCOP] d1phk_ 5.1.1.1.6 gamma-subunit of glycogen phosphorylase kinase 2e-23
 [SCOP] d1abo_ 5.1.1.1.14 Protein kinase CK2, alpha subunit [Maize (Zea)] 1e-23
 45 [SCOP] d3lck_ 5.1.1.2.2 Lymphocyte kinase (lck) [Human (Homo sapiens)] 3e-22
 [SCOP] d2erk_ 5.1.1.1.11 MAP kinase Erk2 [Rat (Rattus norvegicus)] 7e-20
 50 [SCOP] d1cdkb_ 5.1.1.1.2 cAMP-dependent PK, catalytic subunit [Comple] 6e-19
 [SCOP] d1hcl_ 5.1.1.1.1 Cyclin-dependent PK [Human (Homo sapiens)] 5e-21
 [EC] 2.7.1.112 Protein-tyrosine kinase 1e-08
 55 [EC] 2.7.1.126 beta-Adrenergic-receptor kinase 2e-08
 [EC] 2.7.1.117 Myosin-light-chain kinase 1e-09
 [EC] 2.7.1.37 Protein kinase 5e-12

[[EC]] 2.7.1.123 Ca²⁺/calmodulin-dependent protein kinase 4e-09
 [[PIRKW]] phosphotransferase 0.0
 [[PIRKW]] nucleus 9e-09
 5 [[PIRKW]] RNA binding 2e-21
 [[PIRKW]] duplication 8e-10
 [[PIRKW]] tandem repeat 4e-09
 [[PIRKW]] zinc 5e-12
 [[PIRKW]] cell cycle control 2e-09
 10 [[PIRKW]] serine/threonine-specific protein kinase 0.0
 [[PIRKW]] transmembrane protein 2e-09
 [[PIRKW]] zinc finger 8e-10
 [[PIRKW]] oncogene 6e-12
 [[PIRKW]] autophosphorylation 0.0
 15 [[PIRKW]] coat protein 1e-11
 [[PIRKW]] magnesium 9e-09
 [[PIRKW]] ATP 0.0
 [[PIRKW]] polyprotein 6e-12
 [[PIRKW]] receptor 9e-09
 20 [[PIRKW]] phosphoprotein 0.0
 [[PIRKW]] sporulation 2e-09
 [[PIRKW]] glycoprotein 9e-09
 [[PIRKW]] growth factor receptor 9e-11
 [[PIRKW]] signal transduction 2e-12
 25 [[PIRKW]] serine/threonine/tyrosine-specific protein kinase
 8e-10
 [[PIRKW]] protein kinase 8e-10
 [[PIRKW]] transforming protein 2e-12
 [[PIRKW]] heme binding 0.0
 30 [[PIRKW]] purine nucleotide binding 2e-10
 [[PIRKW]] calcium binding 4e-09
 [[PIRKW]] meiosis 1e-08
 [[PIRKW]] alternative splicing 1e-11
 [[PIRKW]] P-loop 2e-10
 35 [[PIRKW]] proto-oncogene 2e-12
 [[PIRKW]] segmentation 4e-10
 [[PIRKW]] stress-induced protein 1e-09
 [[PIRKW]] EF hand 4e-09
 [[PIRKW]] cell division 1e-09
 40 [[PIRKW]] calmodulin binding 4e-09
 [[SUPFAM]] LIM protein kinase 8e-10
 [[SUPFAM]] calcium-dependent protein kinase 4e-09
 [[SUPFAM]] rat protein kinase raf 5e-12
 [[SUPFAM]] AMP-activated protein kinase 2e-08
 45 [[SUPFAM]] protein kinase byr2 5e-09
 [[SUPFAM]] SH2 homology 1e-08
 [[SUPFAM]] unassigned Ser/Thr or Tyr-specific protein kinases 0.0
 [[SUPFAM]] leucine-rich alpha-2-glycoprotein repeat homology 9e-09
 50 [[SUPFAM]] double-stranded RNA-binding repeat homology 2e-21
 [[SUPFAM]] histidine--tRNA ligase homology 6e-42
 [[SUPFAM]] SAM homology 5e-09
 [[SUPFAM]] avian retrovirus IC10 gag-Rml-env polyprotein 1e-11
 [[SUPFAM]] LIM metal-binding repeat homology 8e-10
 55 [[SUPFAM]] GCN2 protein 6e-42
 [[SUPFAM]] protein kinase homology 0.0
 [[SUPFAM]] protein kinase C zinc-binding repeat homology 2e-12
 [[SUPFAM]] Ca²⁺/calmodulin-dependent protein kinase II 4e-08

[SUPFAM] beta-adrenergic-receptor kinase 2e-08
 [SUPFAM] kinase-related transforming protein 1e-12
 [SUPFAM] protein kinase A-raf 2e-12
 [SUPFAM] SH3 homology 1e-08
 5 [SUPFAM] Ca2+/calmodulin-dependent protein kinase 4e-09
 [SUPFAM] protein kinase Xa21 9e-09
 [SUPFAM] calmodulin repeat homology 4e-09
 [SUPFAM] protein kinase DUN1 9e-09
 [SUPFAM] pleckstrin repeat homology 9e-09
 10 [SUPFAM] protein kinase TIK 2e-21
 [SUPFAM] protein-tyrosine kinase tec 1e-08
 [SUPFAM] kinase interaction domain homology 9e-09
 [PROSITE] PROTEIN_KINASE_ATP 2
 [PROSITE] PROTEIN_KINASE_ST 1
 15 [PFAM] Eukaryotic protein kinase domain
 [KW] Irregular
 [KW] 3D
 [KW] LOW_COMPLEXITY 10.99 %
 [KW] COILED_COIL 5.26 %
 20
 SEQ AVLGWPAGWAAARARPAMQGGNSGVRKREEEGDGAGAVAAPPAIDFPAEGPDPEYDESDV
 SEG
 COILS
 25
 1jstA

 SEQ PAEIQVLKEPLQPTFPFAVANQLLLVSLLEHLSHVHEPNPLRSRQVFKLLCQTFIKMGL
 30 SEG
 COILS

 1jstA

 35 SEQ LSSFTCSDEFSSLRLHHNRAITHLMRSKERVVRQDPCEDISRIQKIRSREVALEAQTSTRY
 SEG
 COILS

 40 1jstA

 SEQ LNEFEELAILGKGGYGRVYKVRNKLDGQYYAIKKILIKGATKTVCMKVLREVKVLAGLQH
 45 SEG
 COILS

 1jstA
 TTTEEEEEECBTTBCEEEEEETTTTCEEEEEECCTTTTTTTHHHHHHHHHHTTTB
 50 SEQ PNIVGYHTAWIEHVHVIQPRDRAAIELPSLEVLSDQEEEDREQCGVKNDSSSSSIIFAEP
 SEG
 COILS

 1jstA
 55 TTBC.....
 SEQ TPEKEKRFGESDTENQNNKSVKYTTNLVIRESGELESTLELQENGLAGLSASSIVEQQLP
 SEG

COILS

ljstA

5

SEQ LRRNSHLEESFTSTEESEENVNVLGQTEAQYHMLMLHIQMQLCELSLWDWIVERNKRGRE
 SEGXXXXXXXXXXXXXXXXX.....
 COILS

10

ljstA

SEQ YVDESACPYVMANVATKIFQELVEGVFYIHNMGIVHRDLKPRNIFLHGPDAQVKIGDFGL
 SEG
 COILS

15

ljstA

20

SEQ ACTDILQKNTDWTNRNGKRTPHTSRVGTCLYASPEQLEGSEYDAKSDMYSLGVVLELF
 SEG
 COILS

25

ljstA

SEQ QPFGTEMERAEVLTGLRTGQLPESLRKRCVPQAKYIQHLTRRNSSQRPQAIQLLQSELFQ
 SEG
 COILS

30

ljstA

SEQ NSGNVNLTLMKIIIEQEKEIAELKKQLNLLSQDKGVRDDGKDGGVG
 SEGXXXXXXXXXXXXXXXXX
 COILS ..CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC.....
 ljstA

35

40

Prosite for DKFZphamy2_lil.1

PS00107	190->214	PROTEIN_KINASE_ATP	PD0C00100
PS00107	190->215	PROTEIN_KINASE_ATP	PD0C00100
45 PS00108	454->467	PROTEIN_KINASE_ST	PD0C00100

Pfam for DKFZphamy2_lil.1

50

HMM_NAME Eukaryotic protein kinase domain

HMM

55

*YeigRiIGeGsFGtVYkCiWr.TGeIVAIIK.krsms.....FIREI
 +E + I+G+G++G+VYK++++ +G+ +AIK+I K ++
 +LRE+

Query 184
FEELAILGKGGYGRVYKVRNKLDGQYYAIKKILIKGATKTVCMKVLREV 232

5 HMM qIMRrLnHPNIIRFYDwFedddDHI*
++++ L+HPNI+ + +++ ++ H+
Query 233 KVLAGLQHPNIVGYHTAWI-EHVHV 256

10 HMM *IYMIMEYMeGGDLFDYIrrng.....pMsEweIrfIMyQIL
+++ M+++E +L+D+I++++ ++ + + +I+
+++
Query 396 LHIQMQLCCL-
SLWDWIVERNKRGREYVDESACPYVMANVATKIFQELV 443

15 HMM rGMeYLHSMgIIHRDLKPENILIDeN.gqIKIcDFGLARqMn.....
+G+ Y+H+MGI+HRDLKP+NI++ + Q+KI+DFGLA+
Query 444
EGVFYIHNMGIVHRDLKPRNIFLHGPDAQVKIGDFGLACTDILQKNTDWT 493

20 HMMnYerMttfCGTPWYMMAPEVImgnyYttkVDMWSFGCILWEMMT
+ T+++GT Y +PE ++G+++Y+ K+DM+S+G++L
E++
25 Query 494 NRNGKRTPTHTSRVGTCLYA-SPEQ-
LEGSEYDAKSDMYSLGVLLELF- 540

30 HMM GepPFyd..dnMemImrIiqr.frrpfWpnCSeElyDFMrwCWnyDPekR
+PF ++ E + ++ + ++ ++ +C+ +++ + + +++
++R
Query 541 --QPFGTEMERAEVLTGLRTGQLPESLRKRCPVQAKYIQ-
HLTRRNSSQR 587

35 HMM PTFrQILnHPWF*
P++ Q+L++ F
Query 588 PSAIQLLQSELF 599

DKFZphamy2_lil4

5 group: transmembrane proteins

DKFZphamy2_lil4 encodes a novel 617 amino acid protein with similarity to the human l(3)mbt protein homolog.

10 Mutations of the Drosophila l(3)mbt gene lead to malignant brain tumors. The novel protein contains 1 transmembrane domain. No informative BLAST results; No predictive prosite, pfam or SCOP motive

15 The new protein can find application in studying the expression profile of oncogenes and amygdala-specific genes and as a new marker for amygdala cells.

20 similarity to Human l(3)mbt protein homolog mRNA

> 14 exons (HS756623 (EMBLNEW)).

Pedant: TRANSMEMBRANE 1

25 Sequenced by MediGenomix

Locus: /map="22q13.31-13.33"

Insert length: 3071 bp

30 Poly A stretch at pos. 3052, no polyadenylation signal found

```

      1 GGCAGGCCAA TATGGCTTCC TGCACCTGGT GACGCTTGGC GAAACTGAGG
    51 TCTCATGGAG AAGCCCCGGA GTATTGAGGA GACCCCATCT TCAGAACCAA
  35 101 TGGAGGAAGA GGAAGATGAC GACTTGAGGC TGTTTGGTGG CTATGATAGT
    151 TTCCGGAGTT ATAACAGCAG TGTGGGCACT GAGAGCAGCT CCTATCTGGA
    201 GGAGTCAAGT GAAGCAGAAA ATGAGGATCG GGAAGCAGGG GAACTGCCGA
    251 CCTCCCCGCT GCATTTGCTC AGCCCTGGGA CTCCTCGCTC CTGGGATGGC
    301 AGTGGTTCTG AGCCAGCTGT CTGTGAGATG TGTGGTATCG TGGGTACAAG
  40 351 GGAAGCCTTC TTCTCCAAGA CCAAGAGGTT CTGCAGCGTC TCCTGCTCCA
    401 GGAGCTACTC CTCCAACCTC AAGAAAGCCA GTATCTTGGC TAGATTACAG
    451 GGAAAACCAC CGACCAAAAA AGCCAAAGTC CTGCACAAGG CTGCCTGGTC
    501 TGCCAAAATT GGAGCCTTCC TCCACTCTCA AGGGACAGGA CAGCTGGCAG
    551 ATGGGACACC AACAGGACAA GACGCTCTGG TCTTGGGCTT CGACTGGGGG
  45 601 AAGTTCCTGA AGGATCACAG TTACAAGGCT GCTCCCGTCA GCTGTTTCAA
    651 GCACGTCCCA CTCTATGACC AGTGGGAGGA TGTGATGAAA GGGATGAAGG
    701 TGGAGGTGCT CAACAGTGAT GCTGTGCTCC CCAGCCGGGT GTACTGGATC
    751 GCCTCTGTCA TCCAGACAGC AGGGTATCGG GTGCTGCTTC GGTATGAAGG
    801 CTTTGAAAAT GACGCCAGCC ATGACTTCTG GTGCAACCTG GGAACAGTGG
  50 851 ATGTCCACCC CATTGGCTGG TGTGCCATCA ACAGCAAGAT CCTAGTGCCC
    901 CCACGGACCA TCCATGCCAA GTTCACCGAC TGGAAAGGGCT ACCTCATGAA
    951 ACGGCTGGTG GGCTCCAGGA CGCTTCCCGT GGATTTCAC ATCAAGATGG
   1001 TGGAGAGCAT GAAGTACCCC TTTAGGCAGG GCATGCGGCT GGAAGTGGTG
   1051 GACAAGTCCC AGGTGTCACG CACTCGCATG GCTGTGGTGG ACACAGTAAT
  55 1101 CGGGGGTTCG CTACGGCTCC TCTACGAGGA TGGTGACAGT GACGACGACT
   1151 TCTGGTGCCA CATGTGGAGC CCCCTGATCC ACCCAGTGGG TTGGTCAAGA
   1201 CGTGTGGGCC ACGGCATCAA GATGTCAGAG AGGCGAAGTG ACATGGCCCA
   1251 TCACCCACCC TTCCGGAAGA TCTACTGTGA TGCCGTTCCT TACCTCTTCA

```

```

1301 AGAAGGTACG AGCAGTCTAC ACAGAAGGCG GTTGGTTTGA GGAAGGGATG
1351 AAGCTGGAGG CCATTGACCC CCTGAATCTG GGCAACATCT GCGTGGCAAC
1401 TGTCTGTAAG GTTCTCCTGG ATGGATACCT GATGATCTGT GTGGACGGGG
1451 GGCCCTCCAC AGATGGCTTG GACTGGTTCT GCTACCATGC CTCTTCCCAC
5 1501 GCCATCTTCC CGGCCACCTT CTGTCAGAAG AATGACATTG AGCTCACACC
1551 GCCAAAAGGT TATGAGGCAC AGACTTTCAA CTGGGAGAAC TACTTGGAGA
1601 AGACCAAGTC GAAAGCCGCT CCATCGAGAC TCTTTAACAT GGATTGCCCCA
1651 AACCATGGCT TCAAGGTGGG CATGAAGCTG GAGGCCGTGG ACCTGATGGA
1701 GCCCCGGCTC ATCTGTGTGG CCACGGTGAA ACGAGTGGTG CATCGGCTCC
10 1751 TCAGCATCCA CTTTGACGGC TGGGACAGCG AGTACGACCA GTGGGTGGAC
1801 TGCGAGTCCC CAGACATCTA CCCCCTCGGC TGGTGTGAGC TCACCGGCTA
1851 CCAGCTCCAC CCTCCTGTGG CCGCAGGTGT GGGCTCTCGT GGGCCTAAGA
1901 GGCTCTGACT TTCTTTCTTC TTCTTTTTC CTCTTCCCC CGCCCTGTG
1951 CCCATCTCCG TTCTTTGGCA TGAGGTGGAG ATGTCTCATG GACCACTTTA
15 2001 AGTAGAGAGT GAGCCCCGTC ACCCAGCCCC TGCTCCTGAC TTCTCTGTCT
2051 CCCTTTCCCT CTGGCCTGCA GAGCTCCTTC CTTCATCTTG CCCACTCTGT
2101 CATATGTTCTG TGCCCTTGTG CACCCAGGTA AACTACCCAG GTCCCTCTGA
2151 GCAGCCCTGG TAACAAGGGT GGGGAAGAAG GACAGCTGTT CTCCGGCCCCC
2201 TCCTCCAGCC CCGCCCTCTC CTCATTGCCC AGGTTTGGCT TCCTGTCTTG
20 2251 GGGTGTCTCG TGTGGGAGGG TGGATGGGGT CTCGGGATGC GCCTGTGCCC
2301 TGTGTCTCTC CAGGGACCCT CTTCTCATCT CTTTCACCC TGTCTTTCAA
2351 CAACAGAACC GGCCACACCG CTGAAGGCCA AAGAGGCCAC AAAGAAGAAA
2401 AAGAAACAGT TTGGGAAGAA AAGGAAAAGA ATCCCGCCCA CTAAGACGCG
2451 ACCCTCAGA CAGGGGTCCA AGAAGCCCT GCTGGAGGAC GACCTCAGG
25 2501 GTGCCAGGAA GATCTCGTCG GAGCCTGTT CTGGCGAGAT CATTGCTGTG
2551 CGTGTGAAGG AAGAGCATCT AGACGTGGCC TCGCCCGACA AGGCTTCAAG
2601 TCCAGAGCTG CCTGTCTCCG TCGAGAACAT CAAGCAGGAA ACAGACGACT
2651 GAGCCTTCTT GCCTCCAGCC TGGCTTCTAG CTGGAAGCCA GCCCAGCGTT
2701 TCTCTACCAC CACCACCTG CTCCACCTG ACTTTGGCTT GGAGACTGAT
30 2751 CCTCTCTGTG TAAATTCTGC CCGGTGCTGT GAAGGCTGGA CGGTGGAGGA
2801 CCTGCTGGGG TCTCCTGGGA CCCGCTGTT GCTTCTGCCC TCCCCTGTGG
2851 AAAGGTCTAT ATGACGGGCC GCCTGAGGCC CCAGAACTCG TCTGTGAACC
2901 ACCTTTTCCA GCCAGAGTTC CCAAAGCTGG AACGCTAGCT GCCTGCTCTT
2951 CCTTAAGATG GCCTCCCCC GACCCGCCAC GGCCCTCAGT TGCCAGGGAT
35 3001 GGGGCCACCA CTGTCACACT GTGGAATACA AGACAGTGAA CTCTGTCTGC
3051 CTAATAAAAAA AAAAAAAAAA A

```

BLAST Results

40

Entry HS756G23 from database EMBLNEW:
 Human DNA sequence from clone 756G23 on chromosome 22q13.31-13.33
 Score = 3939, P = 0.0e+00, identities = 875/954

45

Entry U89358_1 from database TREMBL:
 product: "1(3)mbt protein homolog"; Human 1(3)mbt protein
 homolog
 mRNA, complete cds.
 50 Score = 505, P = 7.2e-45, identities = 123/320, positives =
 170/320,
 frame +1

55

Entry AB014581_1 from database TREMBL:
 gene: "KIAA0681"; product: "KIAA0681 protein"; Homo sapiens
 mRNA for
 KIAA0681 protein, partial cds.

Score = 503, P = 1.4e-46, identities = 122/307, positives = 163/307,
frame +1

5

Medline entries

10 No Medline entry

Peptide information for frame 1

15

ORF from 55 bp to 1905 bp; peptide length: 617
Category: similarity to known protein
Classification: unclassified

20

1 MEKPRSI EET PSSEPM EEEE DDDLELF GGY DSFRSYNSSV GSESSSYLEE
51 SSEAENEDRE AGELPTSPLH LLSPGTPRSL DSGSGSEPAVC EMCGIVGTRE
101 AFFSKTKRFC SVSCSRYSYSS NSKKASILAR LQGKPPTKKA KVLHKAWSA
151 KIGAF LHSQG TGQLADGTPT GQDALVLGFD WQKFLKDHSY KAAPVSCFKH
25 201 VPLYDQWEDV MKGMKVEVLN SDAVLPSRVY WIASVIQTAG YRVLLRYEGF
251 ENDASHDFWC NLGTVDVHPI GWCAINSKIL VPPRTIHAKF TDWKG YLMKR
301 LVGSRTL PVD FHIKMVESMK YPFRQGMRL E VVDKSQVSR T RMAVVDTVIG
351 GRLRLLYEDG DSDDDFWCHM WSPLIHPVGW SRRVGHG IKM SERRSDMAHH
401 PTFRKIYCDA VPYLFKKVRA VYTEGGWFEE GMKLEAIDPL NLGNICVATV
30 451 CKVLLDGYLM ICVDGGPSTD GLDWFCYHAS SHAI F PATFC QKNDIELTPP
501 KGYEAQTFNW ENYLEKTKSK AAPSR LFNMD CPNHGFKVGM KLEAVDLM EP
551 RLICVATVKR VVHRLLSIHF DGWDSEYDQW VDCESPDIYP VGWCELTGYQ
601 LQPPVAAGVG SRGPKRL

35

BLASTP hits

No BLASTP hits available

40

Alert BLASTP hits for DKFZphamy2_1114, frame 1

TREMBL:AB014581_1 gene: "KIAA0681"; product: "KIAA0681 protein";
Homo

45 sapiens mRNA for KIAA0681 protein, partial cds., N = 1, Score =
503, P
= 3.9e-48

50 TREMBL:U89358_1 product: "1(3)mbt protein homolog"; Human
1(3)mbt

protein homolog mRNA, complete cds., N = 1, Score = 505, P =
6.2e-48

55 >TREMBL:U89358_1 product: "1(3)mbt protein homolog"; Human
1(3)mbt protein
homolog mRNA, complete cds.
Length = 772

HSPs:

Score = 505 (75.8 bits), Expect = 6.2e-48, P = 6.2e-48

5 Identities = 123/313 (39%), Positives = 170/313 (54%)

Query: 293 WKG YLMKRLVGSRTLPVDFH--

IKMVESMKYPFRQGMRLVVVDKSQVSRTRMAVVDTVIG 350

W+ YL ++ + T PV + V K F+ GM+LE +D S +

10 V V G

Sbjct: 208 WESYLEEQK--

AITAPVSLFQDSQAVTHNKNKGFKLGMKLEGIDPQHPSMYFILTVAEVC 265

Query: 351 GRLRLLYEDGDSDDDFWCHMWSPLIHPVGWSRRVGHGKIMSE--

15 RRSDMAHHPTFRKIY 407

RLRL + DG S+ DFW + SP IHP GW + GH +++ + + +

Sbjct: 266 YRLRLHF-

DGYSECHDFWVNANSPDIHPAGWFECTGHKLQLPKGYKEEFSSWSQYMCSTR 324

20 Query: 408 CDAVP-

YLFKKVRAVYTEGGWFEEGMKLEAIDPLNLGNICVATVCKVLLDGYLMICVDGG 466

A P ++F G F+ GMKLEA+D +N +CVA+V V+ D

++ D

Sbjct: 325 AQAAPKHMVSQSHSPPLG-FQVGMKLEAVDRMNPSLVCVASVTDVV-

25 DSRFLVHFDNW 382

Query: 467 PSTDGLDWFYHASSHAIFPATFCQKNDIELTPPKGY-

EAQTFNWNYLEKTKSKAAPSR 525

T D++C SS I P +CQK LTPP+ Y + F WE YLE+T

30 + A P+

Sbjct: 383 DDT--YDYWC-

DPSSPYIHPVGWCQKQKPLTPPQDYPDPDNFCWEKYLEETGASAVPTW 439

Query: 526

35 LFNMDCPNHGFKVGMKLEAVDLMEPRLICVATVKRVVHRLLSIHFDGWDSEYDQWVDCES 585

F + P H F V MKLEAVD P LI VA+V+ V + IHFDGW

YD W+D +

Sbjct: 440 AFKVR-

PPHSFLVNMKLEAVDRRNPALIRVASVEDVEDHRIKIHFDGWSHG YDFWIDADH 498

40

Query: 586 PDIYPVGWCEL TGYQLQPPV 605

PDI+P GWC TG+ LQPP+

Sbjct: 499 PDIHPAGWCSKTGHPLQPP 518

45 Score = 333 (50.0 bits), Expect = 4.1e-27, P = 4.1e-27

Identities = 103/324 (31%), Positives = 151/324 (46%)

Query: 179 FDWKGFLKDHSYKAAPVSCFKHVPLYDQWEDVMK-

GMKVEVLNSDAVLPSRVYWIASVIQ 237

50 + W +L++ APVS F+ ++ K GMK+E + D PS

+Y+I +V +

Sbjct: 206 WSWESYLEEQKAITAPVSLFQDSQAVTHNKNKGFKLGMKLEGI--DPQHPS-

MYFILTVAE 262

55 Query: 238

TAGYRVLLRYEGFENDASHDFWCNLTVDVHPIGWCAINSKILVPPRTIHAKFTDWKGYL 297

GYR+ L ++G+ HDFW N + D+HP GW L P+ +

W Y+

Sbjct: 263 VCGYRLRLHFDGYSE--
CHDFWVNANSPDIHPAGWFEKTGHKLQLPKGYKEEEFSWSQYM 320

Query: 298 MKRLVGSRTLPLVDFHIKMOVESMKYP---

5 FRQGMRLLEVVDKSQVSRTRMAVVDTVIGGRLR 354
+R H+ + +S P F+ GM+LE VD+ S +A V
V+ R

Sbjct: 321 CS----
TRAQAAPKHMVFVSQSHSPPLGFQVGMKLEAVDRMNPSLVCVASVTDVVDSRFL 376

10 Query: 355 LLYEDGDSDDDFWCHMWSPLIHPVGWSRRVGHGIKMSERRSD---
MAHHPTFRKIYCDAY 411

+ + + + D D+WC SP IHPVGW ++ G + + D
+ AV

15 Sbjct: 377
VHFDNWDYDYWCDPSSPYIHPVGWCQKQKPLTPPDYDPDNFCWEKYLEETGASAV 436

20 Query: 412
PYLFKKVRAVYTEGGWFEEGMKLEAIDPLNLGNICVATVCKVLLDGYLMICVDGGPSTDG 471
P KVR ++ F MKLEA+D N I VA+V V D + I

DG + G
Sbjct: 437 PTWAFKVRPPHS-----FLVNMKLEAVDRRNPAIRVASVEDVE-
DHRIKIHFDDGW--SHG 489

25 Query: 472 LDWFCYHASSHAIFPATFCQKNDIELTPPKG 502
D F A I PA +C K L PP G

Sbjct: 490 YD-FWIDADHPDIHPAGWCSKTGHPLQPPLG 519

30 Score = 236 (35.4 bits), Expect = 2.5e-16, P = 2.5e-16
Identities = 47/110 (42%), Positives = 66/110 (60%)

Query: 499 PPKGYEAQTFNWNENYLEKTKSKAAPSRLE-NMDCPNH---
GFKVGMKLEAVDLMEPRLIC 554

P G + + ++WE+YLE+ K+ AP LF + H GFK+GMKLE +D
35 P +

Sbjct: 197
PATGEKKECWSWESYLEEQKAITAPVSLFQDSQAVTHNKNKGFKLGMKLEGIDPQHPSMYF 256

40 Query: 555 VATVKRVVHRLLSIHFDGWDSEYDQWVDCESPDIYPVGWCELTGYQLQPP
604

+ TV V L +HFDG+ +D WV+ SPDI+P GW E TG++LQ P
Sbjct: 257 ILTVAEVCYRLRLHFDGYSECHDFWVNANSPDIHPAGWFEKTGHKLQLP
306

45 Pedant information for DKFZphamy2_1114, frame 1

50 Report for DKFZphamy2_1114.1

[[LENGTH]] 617

[[MW]] 69264.11

[[pI]] 6.05

55 [[HOMOL]] TREMBL:U89358_1 product: "1(3)mbt protein
homolog"; Human 1(3)mbt protein homolog mRNA, complete cds. 1e-47

[[BLOCKS]] BL01206A Amiloride-sensitive sodium channels proteins

[[KW]] TRANSMEMBRANE 1
[[KW]] LOW_COMPLEXITY 9.40 %

5 SEQ MEKPRSIEETPSSEPMEEEEDDDLELFGGYDSFRSYNSSVGSESSSYLEESSEAEENEDRE
SEGxx
PRD cccccceeecc
MEM
10 SEQ AGELPTSPLHLLSPGTPRSLDGSSEPAVCEMCGIVGTREAFFSKTKRFCVSCSRSYSS
SEGxxxxxxxxxxxx
PRD cccccccccccccccccccccccccccccceeecccccccccccccccccccccccccccccc
MEM
15 SEQ NSKKASILARLQGKPPTKKAKVLHKAAWSAKIGAF LHSQGTGQLADGTPTGQDALVLGFD
SEG xxxxxx.....
PRD ccchhhhhhhhhccccccccchhhhhhhhhhhhhhhhhccccccccccccccccccccceeecc
MEM
20 SEQ W GKFLKDHSYKAAPVSCFKHVPLYDQWEDVMKGMKVEVLNSDAVLPSRVYWIASVIQTAG
SEG
PRD chhhhhhccccccccccccccccccccccccchhhhhhheeeccccccccceeehhhhhhhhhhc
MEM
25 SEQ YRVLLRYEGFENDASHDFWCNLGTVDVHPIGWCAINSKILVPPRTIHAKFTDWKGYLMKR
SEG
PRD ceeeeeeccchhhhhhh
MEM
30 SEQ LVGSRTL PVD FHIKMVESMKYPFRQGMRLLEVVDKSQVSRTRMAVVDTVIGGRLRLLYEDG
SEG
PRD hcc
MEM
35 SEQ DSDDDFWCHMWSPLIHPVGWSRRVGHGIKMSERRSDMAHHPTFRKIYCDAVPYLFKKVRA
SEG
PRD cccccceeeccchhhhhhcccccccccccccc
MEM
40 SEQ VYTEGGWFEEGMKLEAIDPLNLGNICVATVCKVLLDG YLMICVDGGPSTDGLDWFCYHAS
SEG
PRD cccccchhhhhhheeecc
MEMMMMMMMMMMMMMMMMMMMMM.....
45 SEQ SHAIFPATFCQKNDIELTPPKG YEAQTFNWENYLEKTKSKAAPSR LFNMDCPNHGFKVGM
SEG
PRD cchhhhhhhhhhhccccccccccccccccchhhhhhe
MEM
50 SEQ KLEAVDLMEPR LICVATVKRVVHRLLSIHFDGWDSEYDQWVDCESPDIYPVGWCELTGYQ
SEG
PRD eecc
MEM
55 SEQ LQPPVAAGVGSRGPKRL
SEG
PRD ccccccccccccccccccc
MEM

(No Prosite data available for DKFZphamy2_l1l4.1)

5 (No Pfam data available for DKFZphamy2_l1l4.1)

DKFZphamy2_li24

5 group: differentiation/development

DKFZphamy2_li24 encodes a novel 835 amino acid protein without partial similarity to rattus norvegicus Notch2 protein.

10 Notch family molecules are thought to be negative regulators of neuronal differentiation in early brain development. Notch2 is expressed not only by neuronal cells in the embryonic brain, but also by glial cells in the postnatal brain. The new protein represents a new member of this family and may be involved in
15 specific differentiation or developmental pathways of the nervous system.

The new protein can find application in modulating development and differentiation of amygdala cells.

20

putative protein

probably complete cds.

25

Sequenced by MediGenomix

Locus: unknown

30 Insert length: 2768 bp

Poly A stretch at pos. 2714, polyadenylation signal at pos. 2697

```

35      1 AGAAATCTTC AGCCAAACAG CTGCAGGAAG TAGAGAAGGT TAAACCCAG
      51 AGTGAGAAAG TTCATCAGAC TCTGATTCTG GACCCAGCAC AGAGGAAGAG
      101 ACTCCAGCAG CAGATGCAGC AGCACGTTCA GCTCTTGACC CAAATCCACC
      151 TTCTTGCCAC CTGCAACCCC AACCTCAATC CGGAGGCCAC TACCACCAGG
      201 ATATTTCTTA AAGAGCTGGG AACCTTTGCT CAAAGCTCCA TCGCCCTTCA
      251 CCATCAGTAC AACCCCAAGT TTCAGACCCT GTTCCAACCC GTAACCTTGA
40      301 TGGGAGCTAT GCAGCTGATT GAAGACTTCA GCACACATGT CAGCATTGAC
      351 TGCAGCCCTC ATAAAACTGT CAAGAAGACT GCGAATGAAT TTCCCTGTTT
      401 GCCAAAGCAA GTGGCTTGGA TTCTGGCCAC AAGCAAGGTT TTCATGTATC
      451 CAGAGTTACT TCCAGTGTGT TCCCTGAAGG CAAAGAATCC CCAGGATAAG
      501 ATCGTCTTCA CCAAGGCTGA GGACAATTTG TTAGCTTTAG GACTGAAGCA
45      551 TTTTGAAGGA ACTGAGTTTC CTAATCCTCT AATCAGCAAG TACCTTCTAA
      601 CCTGCAAAAC TGCCCACCAA CTGACAGTGA GAATCAAGAA CCTCAACATG
      651 AACAGAGCTC CTGACAACAT CATTAAATTT TATAAGAAGA CCAAACAGCT
      701 GCCAGTCCTA GGAAAATGCT GTGAAGAGAT CCAGCCACAT CAGTGGGAAGC
      751 CACCTATAGA GAGAGAAGAA CACCGGCTCC CATTCTGGTT AAAGGCCAGT
50      801 CTGCCATCCA TCCAGGAAGA ACTGCGGCAC ATGGCTGATG GTGCTAGAGA
      851 GGTAGGAAAT ATGACTGGAA CCACTGAGAT CAACTCAGAT CGAAGCCTAG
      901 AAAAAGACAA TTTGGAGTTG GGGAGTGAAT CTCGGTACCC ACTGCTATTG
      951 CCTAAGGGTG TAGTCCTGAA ACTGAAGCCA GTTGCCACCC GTTTCCCCAG
      1001 GAAGGCTTGG AGACAGAAGC GTTCATCAGT CCTGAAGCCC CTCCTTATCC
55      1051 AACCACGCC CTCTCTCCAG CCCAGCTTCA ACCCTGGGAA AACACCAGCC
      1101 CGATCAACTC ATTCAGAAAG CCCTCCGAGC AAAATGGTGC TCCGGATTCC
      1151 TCACCCAATA CAGCCAGCCA CTGTTTTACA GACAGTTCCA GGTGTCCCTC
      1201 CACTGGGGGT CAGTGGAGGT GAGAGTTTTG AGTCTCCTGC AGCACTGCCT
```

1251 GCTGTGCCCC CTGAGGCCAG GACAAGCTTC CCTCTGTCTG AGTCCCAGAC
 1301 TTTGCTCTCT TCTGCCCCCTG TGCCCAAGGT AATGCTGCCC TCCCTTGCCC
 1351 CTTCTAAGTT TCGAAAGCCA TATGTGAGAC GGAGACCCTC AAAGAGAAGA
 1401 GGAGTCAAGG CCTCTCCCTG TATGAAACCT GCCCTGTTA TCCACCACCC
 5 1451 TGCATCTGTT ATCTTCACTG TTCTTGCTAC CACTGTGAAG ATTGTGAGCC
 1501 TTGGCGGTGG CTGTAACATG ATCCAGCCTG TCAATGCGGC TGTGGCCAG
 1551 AGTCCCCAGA CTATTCCCAT CACTACCCCTC TTGGTTAACC CTACTTCCTT
 1601 CCCCTGTCCA TTGAACCAGT CCCTTGTTGGC CTCCTCTGTC TCACCCTTAA
 1651 TTGTTTCTGG CAATTCTGTG AATCTTCCTA TACCATCCAC CCCTGAAGAT
 10 1701 AAGGCCCCACG TGAATGTGGA CATTGCTTGT GCTGTGGCTG ATGGGGAAAA
 1751 TGCCTTTCAG GGCCTAGAAC CCAAATTAGA GCCCCAGGAA CTATCTCCTC
 1801 TCTCTGCTAC TGTTTTCCCG AAAGTGGAAC ATAGCCCAGG GCCTCCACTA
 1851 GCAGATGCAG AGTGCCAAGA AGGATTGTCA GAGAATAGTG CCTGTCGCTG
 1901 GACCGTTGTG AAAACAGAGG AGGGGAGGCA AGCTCTGGAG CCGCTCCCTC
 15 1951 AGGGCATCCA GGAGTCTCTA AACAACCCCTA CCCCTGGGGA TTTAGAGGAA
 2001 ATTGTCAAGA TGGAACCTGA AGAAGCTAGA GAGGAAATCA GTGGATCCCC
 2051 TGAGCGTGAT ATTTGTGATG ACATCAAAGT GGAACATGCT GTGGAATTGG
 2101 ACACTGGTGC CCCAAGCGAG GAGTTGAGCA GTGCTGGAGA AGTAACGAAA
 2151 CAGACAGTCT TACAGAAGGA AGAGGAGAGG AGTCAGCCAA CTAAAACCCC
 20 2201 TTCATCTTCT CAAGAGCCCC CTGATGAAGG AACCTCAGGG ACAGATGTGA
 2251 ACAAAGGATC ATCAAAGAAT GCTTTGTCTT CAATGGATCC TGAAGTGAGG
 2301 CTTAGTAGCC CCCAGGGGAA GCCAGAAGAT TCATCCAGTG TTGATGGTCA
 2351 GTCAGTGGGG ACTCCAGTTG GGCCAGAAAC TGGAGGAGAG AAGAATGGGC
 2401 CAGAAGAAGA GGAAGAAGAG GACTTTGATG ACCTCACCCA AGATGAGGAA
 25 2451 GATGAAATGT CATCAGCTTC TGAGGAATCT GTGCTTTCTG TCCCAGAACT
 2501 CCAGGTGAGA GCTGGAGAAT ATTCTCAAGT ATTTCTGTGA CTCAGTAATA
 2551 TGTATCACTT ATTGATATGC CACCTGCTTG CTTGCTGCAC TATGGATAGT
 2601 CCTAAATCA TTTGTATTTG ATTTGTGAAT GCATTATGGG ACATGATTGT
 2651 GGAGTTGAGG TGAAATGAGA TGGAAAGGAT GAAATTTTAC TTATTATATT
 30 2701 AAACTCGTTT ACACATTAAG AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
 2751 AAAAAAAAAA AAAAAAAAAA

BLAST Results

35

Entry RNN0TCHX from database EMBL:

Rat notch 2 mRNA.

Score = 818, P = 1.6e-26, identities = 216/277

40

Medline entries

45

No Medline entry

50

Peptide information for frame 3

ORF from 114 bp to 2618 bp; peptide length: 835

Category: putative protein

55

Classification: Differentiation/Development

1 MQQHVQLLTQ IHLLATCNPN LNPEATTTRI FLKELGTFAQ SSIALHHQYN
 51 PKFQTLFQPC NLMGAMQLIE DFSTHVSIDC SPHKTVKKA NEFPCLPKQV

5

10

15

20

25

Pedant information for DKFZphamy2_li24, frame 3

30

35

40

45

50

55

-137-

5
SEQ KTVKKTANEFPCLPKQVAWILATSKVFMYPELLPVCSLKAKNPQDKIVFTKAEDNLLALG
SEG
PRD eeeeeccccccccchhhhhhhccceeeccccccccccccccccceeeeeecchhhhhh

10
SEQ LKHFEGETFPNPLISKYLLTCKTAHQTLTVRIKNLNMNRAPDNIKFYKKTQLPVLGKCC
SEG
PRD hheeeccccccccceeeeeeehhhhhhhhheeeccccccccceeeeeecccccccccceee

15
SEQ EEIQPHQWKPIEREEHRLPFWLKASLPSIQEELRHMA DGAREVGNMTGTTEINSDRSLE
SEG
PRD eeeccccccccchhhhhccceeeecchhhhhhhhhhhhhhhhhhhhhccccccccccccceee

20
SEQ KDNLELGSESRYPLLLPKGVVLKLPVATRFPRKAWRQKRSSVLKPLLIQPSPSLQPSFN
SEG
PRD eccccccccccccccccceeeeeeeeeecchhhhhcccccccccccccccccccccccccc

25
SEQ PGKTPARSTHSEAPPSKMVLRIHPHIQPATVLQTVPGVPPLGVSGGESFESPAALPAVPP
SEG
PRD cccccccccccccccccceeeccccccccceeeeeeccccccccccccccccccccccccccc

30
SEQ EARTSFPLSESQTLLSSAPVPKVMLPSLAPSKFRKPYVRRRPSKRRGVKASPCMKPAPVI
SEG
PRD ccc

35
SEQ HHPASVIFTVPATTVKIVSLGGGCNMIQPVNAAVAQSPQTIPITTLLVNPTSFPCLNQS
SEG
PRD cccccceccccccccceeeeecccccccccccccccccccccccccccccccccecccccccccc

40
SEQ LVASSVSPLIVSGNSVNLPIPSTPEDKAHVNVDIACAVADGENAFQGLEPKLEPQELSPL
SEG
PRD ccc

45
SEQ SATVFPKVEHSPGPPLADAECQEGLSSENSACRWTVVKTEEGRALEPLPQGIQESLNNPT
SEG
PRD ccc

50
SEQ PGDLEEIVKMEPEEAREEISGSPERDIDDIKVEHAVELDTGAPSEELSSAGEVTKQTVL
SEG
PRD cccccccccccccccccceeeccccccccccccccccccccccccccccccccccccchhh

55
SEQ QKEEERSQPTKTPSSSQEPPDEGTSGTDVNKGSSKNALSSMDPEVRLSSPPGKPEDSSSV
SEG
PRD hhhhhhcc

60
SEQ DGQSVGTPVGPETGGGKNGPEEEEEEDFDDLTQDEEDEMSSASEESVLSVPELQVRAGEY
SEG
PRD cccccccccccccccccchhhhhhhccchhhhhhhhhhhccccccccccccceeeecccc

65
SEQ SQVFRGLSNMYHLLICHLLACCTMDSPKIICI
SEG
PRD eeeeehhhhhhhhhhhhhhhhhhhhcccccccccc

55 (No Prosite data available for DKFZphamy2_li24.3)

(No Pfam data available for DKFZphamy2_li24.3)

DKFZphamy2_1j19

5

group: differentiation/development

10 DKFZphamy2_1j19 encodes a novel 150 amino acid protein with high similarity to the allograft inflammatory factor-1 of *Cyprinus carpio*.

15 Allograft inflammatory factor-1 (AIF-1) is a protein involved in allograft rejection. In experimental autoimmune encephalomyelitis (EAE), neuritis (EAN) and uveitis (EAU) it is produced by macrophages and microglia cells.

20 The new protein can find clinical application in the development of tools to enhance the compatibility of transplanted tissues as well as in expression profiling of autoimmune diseases and infections.

25 strong similarity to allograft inflammatory factor-1 (*Cyprinus carpio*)

identical to DKFZphamy2_1n1

30 Sequenced by MediGenomix

Locus: /map="504.9 cR from top of Chr9 linkage group"

Insert length: 3381 bp

35 Poly A stretch at pos. 3362, polyadenylation signal at pos. 3344

```

      1 GCCGGAGCCC GGACCAGGCG CCTGTGCCTC CTCCTCGTCC CTCGCCGCGT
      51 CCGCGAAGCC TGGAGCCGGC GGGAGCCCCG CGCTCGCCAT GTCGGGCGAG
40 101 CTCAGCAACA GGTTCGAAGG AGGGAAGGCG TTCGGCTTGC TCAAAGCCCCG
      151 GCAGGAGAGG AGGCTGGCCG AGATCAACCG GGAGTTTCTG TGTGACCAGA
      201 AGTACAGTGA TGAAGAGAAC CTTCCAGAAA AGCTCACAGC CTTCAAAGAG
      251 AAGTACATGG AGTTTGACCT GAACAATGAA GGCGAGATTG ACCTGATGTC
      301 TTAAAGAGG ATGATGGAGA AGCTTGTTGT CCCCAGACC CACCTGGAGA
45 351 TGAAGAAGAT GATCTCAGAG GTGACAGGAG GGGTCAGTGA CACTATATCC
      401 TACCGAGACT TTGTGAACAT GATGCTGGGG AAACGGTCGG CTGTCTCAA
      451 GTTAGTCATG ATGTTTGAAG GAAAAGCCAA CGAGAGCAGC CCAAAGCCAG
      501 TTGGCCCCC TCCAGAGAGA GACATTGCTA GCCTGCCCTG AGGACCCCGC
      551 CTGGACTCCC CAGCCTTCCC ACCCCATACC TCCCTCCCGA TCTTGCTGCC
50 601 CTTCTTGACA CACTGTGATC TCTCTCTCTC TCATTTGTTT GGTCAATTGAG
      651 GGT TTGTTT TGT TTTT CATC AATGTCTTTG TAAAGCACAA ATTATCTGCC
      701 TTAAAGGGGC TCTGGGTCGG GGAATCCTGA GCCTTGGGTC CCCTCCCTCT
      751 CTTCTTCCCT CTTCCCCGC TCCCTGTGCA GAAGGGCTGA TATCAAACCA
      801 AAAACTAGAG GGGGCAGGGC CAGGGCAGGG AGGCTTCCAG CCTGTGTTCC
55 851 CCTCACTTGG AGGAACCAGC ACTCTCCATC CTTTCAGAAA GTCTCCAAGC
      901 CAAGTTCAGG CTCACTGACC TGGCTCTGAC GAGGACCCCA GGCCACTCTG
      951 AGAAGACCTT GGAGTAGGGA CAAGGCTGCA GGGCCTCTTT CGGGTTTCCT
     1001 TGGACAGTGC CATGGTTCCA GTGCTCTGGT GTCACCCAGG ACACAGCCAC

```

```

1051 TCGGGGGCCCC GCTGCCCCAG CTGATCCCCA CTCATTCCAC ACCTCTTCTC
1101 ATCCTCAGTG ATGTGAAGGT GGAAGGAAA GGAGCTTGGC ATTGGGAGCC
1151 CTTCAAGAAG GTACCAGAAG GAACCCCTCA GTCTGTCTCT CTGGCCACAC
1201 CTGTGCAGGC AGCTGAGAGG CAGCGTGCAg CCTACTGTC CTTACTGGG
5 1251 GCAGCAGAGG GCTTCGGAGG CAGAAGTGAG GCCTGGGGTT TGGGGGGAAA
1301 GGTCAgCTCA GTGCTGTTCC ACCTTTTAGG GAGGATACTG AGGGGACCAG
1351 GATGGGAGAA TGAGGAATAC AAGGTTGCTT GTCTGACCCC AATCTGCTTG
1401 AAGCCAAGAC TGAGAAATAC AAGGTTGCTT GTCTGACCCC AATCTGCTTG
1451 AAACCTGACT CTGCTTCTCT CATTTGTCTT CCTACCCTAC TCACATAATT
10 1501 CACTCATTGA CTCACCTATT CACCAGATAT TTATTGACCT GCTATTATAA
1551 GCTTTACATC CTCCCATGTT GTCTGGCAT GTGCAGTATA CACGGTCTAA
1601 CTCATCTCTC CCCAGATCTC TCAGAACCTT GAGCTTGGGA ATTGAACTGG
1651 GGTCACCTGT GTCTTTTCTT ATGGACTCGC AGGATTTTAG AACCTAATG
1701 CACCCTGGAG GGTAGCTGGG CCAGACTTCT CATTTCAcAG GTGAGGAGAC
15 1751 TGGTGCCCCA CAGGGATTAA GTGCCTTGCC CAAGGTCAGG CTTATCTCCA
1801 GAGGGAGGTT CCCTGGACTG GGGCCcAGAT GTTCAGGGAC CCTGCCTACA
1851 CCTCATTTC AGTGTGGGCT GCCTTAGTTA GTTATGAGAA CAGGGAAGGG
1901 CTGGGAAGAG ACAGCTCCA AGGTCAACAC TTGGAGAGGG TTTCACTTGC
1951 TCTGAAGACC CTGGTCCAGG ATTGCCCCCT TCCCATGCCT TCAAGTCAGC
20 2001 ATCAGGCTTA GGGCAAAGAC CAGGCCCTCT AAGCTGCCTC TTGTAATTCA
2051 TGCAGGAAGA TGTCAAAGTC AGCCCCATCT TGGCTGATCA GGGTGTTCAG
2101 CCTTAACCCC ACCTGTGTTC TGAAGTCTCT TACCCTACCT GCTCAGGACT
2151 GAGACAGTTA TTCACTGAAC ATATTTATTA AGCACTTGCT GTAGGCCAAC
2201 AGTTAAGAAT CCAATAATGA AATGGACAGA TTCATGGAAC TTAGAGTCCA
25 2251 ATAGGAAAGT GAGACCCAGA CAATGACAAT GAGATAAATG TTAGGAAGGG
2301 GGAGGTATGG GGTGACTTCC CTGCAGTCCT GGGGGCCTAC ATGGGCCCAA
2351 GACTGGGTGA GAGTCTTGGC AGAGCCTTTG CAACACCTTA AGTGGACAGG
2401 ACTGGGAGGT CTTGGTGGTT GAGGCCAACG TGGGTTCCTT GCGGCTCCTT
2451 AGTCACCTCT GATAGCAGAT TGAGGGAGGA AAACAGGTAA GGCATGAGGA
30 2501 AATGGCCAGG TTGGGTAAAC CCACTGGTTT CAACCAGTTC AGGAATGAGG
2551 TTATTTGGCC ATGACTGGCT GATCTTGAGC TCAAGGATCT GCTTCAAATG
2601 CACACAGGCC TAGTTGAAGT TTAACCCCA GCAAAACATT CCTCCCTGTA
2651 AATGGAAAT CCTACTTCTA CCCCACCCCT GCCCTGTTTT TTGTTTTTTT
2701 TTTCCCCAAG ATCATTAGAT GTCCTCACCC CTCCTCACTG CCTCTCCTCT
35 2751 CTGGGACAGG CTGGGACCTT TGAGGAAGAT AAAGCCTTCC TTGACTACCC
2801 ATCATATTCA GTGTCCCTGT TCCTCACTCA GAGAGGAAGG CAGAACCAGT
2851 CAGGCTTATT TCAGTAAGTT CCACAGTTCT ACAAGACTGC AGGAATTCTC
2901 CTTAAGGGAG GAGAGCAAGC AGGTGTGGCC CCAGCTTCTG GAAATGGCAG
2951 AAGAGAGGGT TTTCTCATTG AATGGGGGTG GGGGCTCGTG TGTCCTGGGA
40 3001 AACCCCATCA GTCCCTTCAT TTCTTGAGAC TCAACTCCTG GGAGGAGAGG
3051 GTCTCAAGAG TTGTCCCTGG AAGGAGGGCG GGGGCAGTCT GCATCTATTT
3101 CAGGTTGTGG CTCTTGTTTC TAGGACTCTT ACTTCTCTGG CTAAGGGCTC
3151 AGCTTCTTGG GACTTCAACC ATCTTCTTTC TGAAAGACCA AATCTAATGT
3201 AACCAgTAAC GTGAGGACTG CCAAGTATGG CTTTGTCCCT ATGACTCAGA
45 3251 GGAGGGTTTG TCGGGCAAAT TCAGGTGGAT GAAGTATGTG TGTCGTGTG
3301 CATGGGAGTG TGCCTGGACT GGGATATCAT CTCTACAGCC TGCAAATAAA
3351 CCAGACAAAC TTAACAAAAA AAAAAAAAAA A

```

50

BLAST Results

Entry AB012309_1 from database TREMBL:

product: "allograft inflammatory factor-1"; Cyprinus carpio mRNA

55

for allograft inflammatory factor-1, complete cds.

Score = 575, P = 3.7e-54, identities = 113/146, positives = 128/146.

frame +2

5 Medline entries

No Medline entry

10

Peptide information for frame 2

15 ORF from 89 bp to 538 bp; peptide length: 150
 Category: strong similarity to known protein
 Classification: unclassified

20 1 MSGELSNRFQ GGAFFGLLKA RQERRLAIEIN REFLCDQKYS DEENLPEKLT
 51 AFKEKYMED LNNEGEIDLM SLKRMMEKLG VPKTHLEMKK MISEVTGGVS
 101 DTISYRDFVN MMLGKRSAVL KLVMMFEGKA NESSPKPVGP PPERDIASLP

25 BLASTP hits

No BLASTP hits available

30 Alert BLASTP hits for DKFZphamy2_1j19, frame 2

No Alert BLASTP hits found

Pedant information for DKFZphamy2_1j19, frame 2

35

Report for DKFZphamy2_1j19.2

40 [LENGTH] 150
 [MW] 17067.86
 [pI] 6.63
 [HOMOL] TREMBL:AB012309_1 product: "allograft inflammatory factor-1"; Cyprinus carpio mRNA for allograft inflammatory factor-1, complete cds. 2e-59
 45 [FUNCAT] 30.04 organization of cytoskeleton [S. cerevisiae, YBR109c] 5e-04
 [FUNCAT] 03.07 pheromone response, mating-type determination, sex-specific proteins [S. cerevisiae, YBR109c] 5e-04
 50 [FUNCAT] 08.19 cellular import [S. cerevisiae, YBR109c] 5e-04
 [FUNCAT] 10.02.99 other morphogenetic activities [S. cerevisiae, YBR109c] 5e-04
 [FUNCAT] 03.22 cell cycle control and mitosis [S. cerevisiae, YBR109c] 5e-04
 55 [FUNCAT] 03.04 budding, cell polarity and filament formation [S. cerevisiae, YBR109c] 5e-04
 [FUNCAT] 03.01 cell growth [S. cerevisiae, YBR109c] 5e-04
 [FUNCAT] 30.05 organization of centrosome [S. cerevisiae, YBR109c] 5e-04

[SCOP] d2mysb_ 1.37.1.5.15 Myosin Essential Chain Myosin
 Regulatory Chain 5e-20
 [SCOP] dlwddb_ 1.37.1.5.14 Myosin Essential Chain Myosin
 Regulatory Chain 3e-05
 5 [SCOP] dlosa_ 1.37.1.5.13 Calmodulin [(Paramecium
 tetraurelia) 3e-16
 [SCOP] dlauib_ 1.37.1.5.19 Calcineurin regulatory subunit
 (B-chain 2e-16
 [PIRKW] duplication 7e-06
 10 [PIRKW] mitosis 7e-06
 [PIRKW] calcium binding 7e-06
 [PIRKW] EF hand 7e-06
 [PIRKW] cell division 7e-06
 [SUPFAM] unassigned calmodulin-related proteins 3e-47
 15 [SUPFAM] calmodulin 7e-06
 [SUPFAM] calmodulin repeat homology 3e-47
 [KW] All_Alpha
 [KW] 3D

20 SEQ MSGELSNRFQGGKAFGLLKARQERRLAIEINREFLCDQKYSDEENLPEKLTAFKEKYMEFD
 1ctr-HHHHHHHHHHHHHT
 25 SEQ LNNEGEIDLMSLKRMMEKLGVPKTHLEMKKMISEVTGGVSDTISYRDFVNMMLGKRSAVL
 1ctr- TTTTTCBCHHHHHHHHHHTTTCCCHHHHHHHHHCTTTTCCCBCHHHHHHHHCCTTTTHH
 SEQ KLVMMFEGKANESSPKPVGPPPERDIASLP
 30 1ctr- HHHHHHTTTTC.....

(No Prosite data available for DKFZphamy2_1j19.2)

35 (No Pfam data available for DKFZphamy2_1j19.2)

DKFZphamy2_24b4

5 group: cell cyle

DKFZphamy2_24b4 encodes a novel 698 amino acid protein with similarity to human STIM1.

10 The stromal interaction molecular 1 gene (STIM1) encodes a type I trans-membrane protein of unknown function, which induces growth arrest and degeneration of the human tumor cell lines G401 and RD but not HBL100 and Calu-6, suggesting a role in the pathogenesis of rhabdomyosarcomas and rhabdoid tumors. There is also strong
15 similarity to a Mus musculus stromal cell protein, which selectively increases interleukin 7-dependent proliferation of pre-B cells. The novel protein contains 1 transmembrane domain.

20 The new protein can find application in modulation of tumour growth.

similarity to STIM1 (Homo sapiens)

25 probably differential polyadenylation: cf. EST-BLAST file.
perhaps complete cds.

Pedant: SIGNAL_PEPTIDE and TRANSMEMBRANE 1

Sequenced by GBF

30

Locus: /map="139.2 cR from top of Chr4 linkage group"

Insert length: 3305 bp

Poly A stretch at pos. 3274, polyadenylation signal at pos. 3260

35

```

1  GGC GCCTTCA TCCG CCTCG ACTCCTGGCC CAGCGTGGGG CTGGCTGCTG
51  CGGCGGCGGC GCTGGGCTGC GTTGCTGGTG CTCGGGCTGC TGGTACCCGG
101 AGCGGCGGAC GGATGCGAGC TTGTGCCCCG GCACCTCCGC GGGCGGCGGG
40 151 CGACTGGCTC TGCCGCAACT GCCGCCTCCT CTCCCGCCGC GGCGGCCGGC
201 GATAGCCCGG CGCTCATGAC AGATCCCTGC ATGTCACTGA GTCCACCATG
251 CTTTACAGAA GAAGACAGAT TTAGTCTGGA AGCTCTTCAA ACAATACATA
301 AACAAATGGA TGATGACAAA GATGGTGGAA TTGAAGTAGA GGAAAGTGAT
351 GAATTCATCA GAGAAGATAT GAAATATAAA GATGCTACTA ATAAACACAG
45 401 CCATCTGCAC AGAGAAGATA AACATATAAC GATTGAGGAT TTATGGAAAC
451 GATGGAAAC ATCAGAAAGT CATAATTGGA CCCTTGAAGA CACTCTTCAG
501 TGGTTGATAG AGTTTGTGTA ACTACCCCAA TATGAGAAGA ATTTTAGAGA
551 CAACAATGTC AAAGGAACGA CACTTCCCAG GATAGCAGTG CACGAACCTT
601 CATTTATGAT CTCCCAGTTG AAAATCAGTG ACCGGAGTCA CAGACAAAAA
50 651 CTTCAGCTCA AGGCATTGGA TGTGGTTTTG TTTGGACCTC TAACACGCCC
701 ACCTCATAAC TGGATGAAAG ATTTTATCCT CACAGTTTCT ATAGTAATTG
751 GTGTTGGAGG CTGCTGGTTT GCTTATACGC AGAATAAGAC ATCAAAAGAA
801 CATGTTGCAA AAATGATGAA AGATTTAGAG AGCTTACAAA GTGCAGAGCA
851 AAGTCTAATG GACTTACAAG AGAGGCTTGA AAAGGCACAG GAAGAAAACA
55 901 GAAATGTTGC TGTAGAAAAG CAAAATTTAG AGCGCAAAAT GATGGATGAA
951 ATCAATTATG CAAAGGAGGA GGCTTGTCGG CTGAGAGAGC TAAGGGAGGG
1001 AGCTGAATGT GAATTGAGTA GACGTCAGTA TGCAGAACAG GAATTGGAAC
1051 AGGTTTCGCAT GGCTCTGAAA AAGGCCGAAA AAGAATTGTA ACTGAGAAGC

```

```

1101 AGTTGGTCTG TTCCAGATGC ACTTCAGAAA TGGCTTCAGT TAACACATGA
1151 AGTAGAAGTG CAATACTACA ATATTA AAAAG ACAAAACGCT GAAATGCAGC
1201 TAGCTATTGC TAAAGATGAG GCAGAAAAAA TTAAAAAGAA GAGAAGCACA
1251 GTCTTTGGGA CTCTGCACGT TGACACACAGC TCCTCCCTAG ATGAGGTAGA
5 1301 CCACAAAATT CTGGAAGCAA AGAAAGCTCT CTCTGAGTTG ACAACTTGTT
1351 TACGAGAACG ACTTTTTCGC TGGCAACAAA TTGAGAAGAT CTGTGGCTTT
1401 CAGATAGCCC ATAACCTCAGG ACTCCCCAGC CTGACCTCTT CCCTTTATTC
1451 TGATCACAGC TGGGTGGTGA TGCCAGAGT CTCCATTCCA CCCTATCCAA
1501 TTGCTGGAGG AGTTGATGAC TTAGATGAAG ACACACCCCC AATAGTGTCA
10 1551 CAATTTCCCG GGACCATGGC TAAACCTCCT GGATCATTAG CCAGAAGCAG
1601 CAGCCTGTGC GCTTCACGCC GCAGCATTGT GCCGTCCTCG CCTCAGCCTC
1651 AGCGAGCTCA GCTTGCTCCA CACGCCCCCC ACCCGTCACA CCTCGGCAC
1701 CCTCACCACC CGCAACACAC ACCACACTCC TTGCCTTCCC CTGATCCAGA
1751 TATCCTCTCA GTGTCAAGTT GCCCTGCGCT TTATCGAAAT GAAGAGGAGG
15 1801 AAGAGGCCAT TTACTTCTCT GCTGAAAAGC AATGGGAAGT GCCAGACACA
1851 GCTTCAGAAT GTGACTCCTT AAATTCTTCC ATTGGAAGGA AACAGTCTCC
1901 TCCTTTAAGC CTCGAGATAT ACCAAACATT ATCTCCGCGA AAGATATCAA
1951 GAGATGAGGT GTCCCTAGAG GATTCTCTCC GAGGGGATTG GCCTGTAAC
2001 GTGGATGTGT CTTGGGGTTC TCCCGACTGT GTAGGTCTGA CAGAACTAA
20 2051 GAGTATGATC TTCAGTCTGT CAAGCAAGT GTACAATGGC ATTTTGGAGA
2101 AATCCTGTAG CATGAACCAG CTTTCCAGTG GCATCCCGGT CCTTAAACCT
2151 CGCCACACAT CATGTTCTCT AGCTGGCAAC GACAGTAAAC CAGTTCAGGA
2201 AGCCCCAAGT GTTGCCAGAA TAAGCAGCAT CCCACATGAC CTTTGTCTA
2251 ATGGAGAGAA AAGCAAAAAG CCATCAAAAA TCAAAAGCCT TTTTAAGAAG
25 2301 AAATCTAAGT GAACTGGCTG ACTTGATGGA ATCATGTTCA AGTGGCATCT
2351 GTAAACTATT ATCCCCACC CTCCACTCCC CACCTTTTTT TTGGTTTAA
2401 TTTAGGAATG TAACTCCATT GGGGCTTTCC AGGCCGGATG CCATAGTGG
2451 ACATCCAGAA GGGCAACTGT CTACTGTCTG CTTATTTAAG TGAATATATA
2501 TAATCAATTC ATCAAGCCAG TTATTACTGA AAAATCATTG AAATGAGACA
30 2551 GTTTACAGTC ATTTCTGCCT ATTTATTTCT GCTTTGTTCT CAGTGATGTA
2601 TATGCAACAT TTTGTTGAAA GCCACGATGG ACTTACAAGC TTTAATGGAC
2651 TCGTAAGCCA GCATGGGCTT GCAAAAATTT CTTGTTTACC AGAGCATCTT
2701 CTTATCTTTC CACAGAGCTA TTTACATCCT GGAATATATA ACTTAAAGA
2751 AGTAAACGT AATTGCACTA CTGTTTTCCA GACTGGAAAA AAAAAAAAT
35 2801 CTCTGCAAGT GAAACTGTAT AGAGTTTATA AAATGACTAT GGATAGGGGA
2851 CTGTTTTCAC TTTTAGATCA AAATGGGTTT TTAAGTAGAA CCTAGGGTTT
2901 CTAATTGACT TGATTTCTGG AAATGAAAAC CCGCGCTTTT ATTATGGGAA
2951 GCTTCTTGAA CTGCATTTAC TATTGTGAAG TTTCAAGTCC CGCTGTAAAG
3001 ATCATGTTGT TTTGTTTTCC CCAGGGCTTT CACTGTGATT TACTGCATTG
40 3051 CAGGCTGTAT GATAAAACAC ACATAATTTA AAGAGAGAAG GCTCTTGATT
3101 CTTATGCAA GTGGAAGAGT TGAAACTTGA TTGAAGGACT TAAACATTTC
3151 ACAACCTTAA GCGGAGGTGG GGGGATATGG GGATTTCAGGC AGTTGTTTAC
3201 ACACTTTGAA TAACTGCAAA GGATTTACGG TTTGTGAAAA ATGTGTACTG
3251 TGGAAAAGAT AATAAATTGA AGACATTA AAAGAAAAA AAAAAAAAAA
45 3301 AAAAA

```

BLAST Results

50

Entry HS5242610_1 from database TREMBL:
 gene: "STIM1"; product: "GOK"; Homo sapiens GOK (STIM1) mRNA,
 complete
 cds.

55

Score = 1397, P = 4.2e-142, identities = 275/447, positives =
 336/447,
 frame +3

Entry MMU47323_1 from database TREMBL:

product: "stromal cell protein"; Mus musculus stromal cell protein

mRNA, complete cds.

5 Score = 1394, P = 8.8e-142, identities = 274/447, positives = 336/447, frame +3

Entry HS917349 from database EMBL:

10 human STS EST167479.

Score = 1390, P = 9.1e-57, identities = 284/287

15

Medline entries

97079692:

20 Parker NJ, Begley CG, Smith PJ, Fox RM.; Molecular cloning of a novel human gene (D11S4896E) at chromosomal region 11p15.5. Genomics 1996 Oct 15;37(2):253-6

96326680:

25 Oritani K, Kincade PW.; Identification of stromal cell products that interact with pre-B cells. J Cell Biol 1996 Aug;134(3):771-82

30

Peptide information for frame 3

35

ORF from 216 bp to 2309 bp; peptide length: 698

Category: similarity to known protein

Classification: Cell signaling/communication

Prosites motifs: RGD (589-591)

40

1 MTDPCMSLSP PCFTEEDRFS LEALQTIHKQ MDDDKDGGIE VEESDEFIRE
 51 DMKYKDATNK HSHLHREDKH ITIEDLWKRW KTSEVHNWTL EDTLQWLIEF
 101 VLPQYKFN RDNNVKGTTL PRIAVHEPSF MISQKISDR SHRQKLQKA
 45 151 LDVVLFGPLT RPPHNWMKDF ILTVSIVIGV GGCWFAYTQN KTSKEHVAKM
 201 MKDLESLQTA EQSLMDLQER LEKAQEEENRN VAVEKQNLER KMMDEINYAK
 251 EEACRLREL R EGAECESRR QYAEQELEQV RMALKKAEKE FELRSSWSVP
 301 DALQKWLQLT HEVEVQYYNI KRQNAEMQLA IAKDEAEKIK KKRSTVFGTL
 351 HVAHSSSLDE VDHKILEAKK ALSELTTCLE ERLFRWQQIE KICGFQIAHN
 50 401 SGLPSLTSSL YSDHSWVVMP RVSIPPYPPIA GGVDLDLDET PPIVSQFPQT
 451 MAKPPGSLAR SSSLCRSRRS IVPSSPQPQR AQLAPHAPHP SHPRHPHPQ
 501 HTPHSLPSPD PDILSVSSCP ALYRNEEEEE AIYFSAEKQW EVPDTASECD
 551 SLNSSIGRKQ SPPLSLEIYQ TSPRKISRQ EVSLEDSSRG DSPVTVDVSW
 601 GSPDCVGLTE TKSMIFSPAS KVYNGILEKS CSMNQLSSGI PVPKPRHTSC
 55 651 SSAGND SKPV QEAPSVARIS SIPHDLCHNG ESKKPSKIK SLFKKKSK

No BLASTP hits available

No Alert BLASTP hits found

Report for DKFZphamy2_24b4.3

MEM

5 SEQ FILTVSIVIGVGGCWFAYTQNKTSKEHVAKMMKDLESLQTAEQSLMDLQERLEKAQEENR
SEG
PRD hhhheeeeeccccceeeccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcc
COILSCC
MEM MMMMMMMMMMMMMMMMM.....

10 SEQ NVAVEKQNLERKMMDEINYAKEEACRLRELREGAECELSRRQYAEQELEQVRMALKKAEK
SEG
PRD ceeeeeHH
COILSCC.....

15 MEM
SEQ EFELRSSWSVPDALQKWLQLTHEVEVQYNIKRQNAEMQLAIKDEAEKIKKKRSTVFGT
SEGXXXXXXXXXXXXXXXX.....
PRD hhhhHCCCCCCHHHHHHHHHHHHHeeeecchhhhhhhhhhhhhhhhhhhhhhhhhhhhhccce
COILS
MEM
SEQ LHVAHSSSLDEV DHKILEAKKALSELTTCLRERLFRWQIEKICGFQIAHNSGLPSLTSS
SEG
25 PRD eeeeeccccchhh
COILS
MEM
SEQ LYS DHSWVVM PRVSIPPYPIAGGVDDLDETPPIVSQFPGTMAKPPGSLARSSSLCRSRR
SEGXXXXXXXXXXXXXXXX.....
PRD CCC
COILS
35 MEM
SEQ SIVPSSPQPQRAQLAPHAPHSPHPRHPHPQHTPHSLPSPDPDILSVSSCPALYRNEEEE
SEG x.....XXXXXXXXXXXXXXXXXXXXXXXXXXXX.....xxxx
PRD eeeccceeeeeeccccchhhhhh
COILS
MEM
SEQ EAIYFSAEKQWEVPDTASECDLNSSIGRKQSPPLSLEIYQTLSPRKISRDEVSLEDSSR
SEG x.....
PRD hhhhhhhhhhhcc
COILS
MEM
50 SEQ GDS PVTVDVSWGSPDCVGLTETKSMIFSPASKVYNGILEKSCSMNQLSSGIPVPKPRHTS
SEG
PRD ccc
COILS
55 MEM
SEQ CSSAGNDSKPVQEA PSVARISSIPHDLCHNGEKS KPSKIKSLFKKKSK

SEGxxxxxxxxxxxxxxxxxxxx
PRD ccc
COILS
MEM

5

Prosites for DKFZphamy2_24b4.3

10 PS00016 660->663 RGD PD0C00016

(No Pfam data available for DKFZphamy2_24b4.3)

DKFZphamy2_24c8

5 group: transmembrane protein

DKFZphamy2_24c8 encodes a novel 454 amino acid protein without similarity to known proteins.

10 The novel protein contains 1 transmembrane region.
No informative BLAST results; No predictive prosite, pfam or SCOP motif.

15 The new protein can find application in studying the expression profile of amygdala-specific genes and as a new marker for amygdala cells.

20 putative protein

EST of GEN-426HD7 is 141 Bp longer at 5'-end perhaps complete cds.

Pedant: TRANSMEMBRANE 1

25 Sequenced by GBF

Locus: /map="b09.7 cR from top of Chr3 linkage group"

30 Insert length: 3200 bp

Poly A stretch at pos. 3177, polyadenylation signal at pos. 3156

```

1  CCTGTCCACA GGGCCCGCTC CAGCAGCCAT GGCAACCACA TCCTCCAAGC
35  51  CAGAGGGGCG CCCTCGAGGG CAGGCTGCCC CCACCATCCT GCTGACAAAG
    101 CCACCGGGGG CCACAGCCG CCCCACCACA GCGCCCCCCC GCACTACCAC
    151 ACGCAGGCCC CCCAGGCCCC CAGGCTCTTC CCGAAAAGGG GCTGGTAATT
    201 CATCACGCCC TGTCCCGCCT GCACCTGGTG GCCACTCCAG GAGTAAAGAA
    251 GGACAGCGAG GACGAAATCC AAGCTCCACA CCTCTGGGGC AGAAGCGGCC
40  301 CCTGGGGAAA ATCTTTCAGA TCTACAAGGG CAACTTCACA GGGTCTGTGG
    351 AACC GGAGCC CTCTACCCTC ACCCCCAGGA CCCCCTCTG GGGCTACTCC
    401 TCTTACCAC AGCCCCAGAC AGTGGCTGCG ACCACAGTGC CCAGCAATAC
    451 CTCATGGGCA CCCACCACCA CCTCCCTGGG GCCTGCAAAG GACAAGCCAG
    501 GCCTTCGCAG AGCAGCCCAG GGGGGTGGTT CTACCTTCAC CAGCCAAGGA
45  551 GGGACACCAG ATGCCACAGC AGCCTCAGGT GCCCCTGTCA GTCCACAAGC
    601 TGCCCCAGTG CCTTCTCAGC GCCCCCACCA CGGTGACCCA CAGGATGGCC
    651 CCAGCCATAG TGACTCTTGG CTTACTGTTA CCCCTGGCAC CAGCAGACCT
    701 CTGTCTACCA GCTCTGGGGT CTTACGGCT GCCACGGGGC CCACCCCAGC
    751 TGCCTTCGAT ACCAGTGTCT CAGCCCCCTC CCAGGGGATT CCTCAGGGAG
50  801 CATCCACAAC CCCACAAGCT CCAACCCATC CCTCCAGGGT CTCAGAAAGC
    851 ACTATTTCTG GAGCCAAGGA GGAGACTGTG GCCACCCTCA CCATGACCGA
    901 CCGGGTGCCC AGTCCTCTCT CCACAGTGGT ATCCACAGCC ACAGGCAATT
    951 TCCTCAACCG CCTGGTCCCC GCGGGGACCT GGAAGCCTGG GACAGCAGGG
55 1001 AACATCTCCC ATGTGGCCGA GGGGGACAAA CCGCAGCACA GAGCCACCAT
    1051 CTGCCTGAGC AAGATGGATA TCGCCTGGGT GATCCTGGCC ATCAGCGTGC
    1101 CCATCTCCTC CTGCTCTGTC CTGCTGACGG TGTGCTGCAT GAAGAGGAAG
    1151 AAGAAGACCG CCAACCCGGA GAACAACCTG AGCTACTGGA ACAACACCAT
    1201 CACCATGGAC TACTTCAACA GGCATGCTGT GGAGCTGCCC AGGGAGATCC

```



```

1251 AGTCCCTTGA AACCTCTGAG GACCAGCTCT CAGAGCCCCG CTCCCCAGCC
1301 AATGGCGACT ATAGAGACAC TGGGATGGTC CTTGTTAACC CCTTCTGTCA
1351 AGAAACACTG TTTGTGGGAA ACGATCAAGT ATCTGAGATC TAACTACAGC
1401 AGGCATCACT TTGCCATTCC GTATTTTTCG TCTCTAAATT ATAAATATAC
5 1451 AAATATATAT ATTATAAATA TAACCTTTGT GTAACCCTGA CTTAATGAGA
1501 AACATTTTCA GCTTTTTTTC CTATGAATTG TCAACATCTT TTTTACAAGT
1551 GTGGTTTAAA AAAAAAAAAA CTTTACAGAA TGATCTGTGG CTTTATAAAA
1601 TAAAGGTATT TCTAAGCAAA GCAGTTGCAT TGATTGCTTC TCTTAATAAC
1651 TATTCTTGAG CACCTGGGGA TCCCAGGAAC CCTGGTCAGG TGAGGTAAGA
10 1701 GACTGACCTC CTGTAGAAGC TGAATGTTAC AGTGGTCAAG CGCACGATTC
1751 TTTGAGTGAT TCTTAAAGCT CTGGTTCCTC TTGATTTGGT GTGACCCCAT
1801 TTCCTCCCTT CTCATACGCA CACCTGTAAA GGGAACTGGA CCGCCTCAGG
1851 GGAAGACGGC AGACTCATGC ACAGAGAAAG AAAAGGGAAC ATCTCATCAC
1901 CTCTGAGGAT GAGTACCTG GAGCCTTATG ACGGCACCAT TGGATGTCTC
15 1951 GTTTAATTCC ATCCAAGTTG TGGATGGCAG GCAGGAGCAT GGAGCCCTCA
2001 GGAATCCATG GAGGACATCA AGGCATCCCA AGGCCATATT CCCCTAACAT
2051 TACTTCCACT GCTAACAACA GGAAGTGCCT TCCCTGGTGG GAAAATGCTC
2101 CCTTTATGCC CATTCCTGTA TCCCCTCCAA CACCCACATC TGCATTAAAC
2151 ACCCGTGCCT TTCTCTTGGG GAGGGTTTAG ATGCAGATCC CGGCCCTGGA
20 2201 GCTTTAAAT GCTTGCCCTT CCTTCTTCAA GGATCAAATG TTTATTGGGG
2251 TTCAGCTTTG TTTTCTCAA AGGCCATGGT ATCGTGCCCC TGAGGAACAT
2301 GTTTATCTAA GAAGCTTTGA GGTAGTAGAG CGATAATTTT TGAAACCTTC
2351 CTCCTGCAAT CTTTAAAAAA GAAAAAAAAG ATTGCCCAA CAAATCATT
2401 GGGAGAAGAC ATCATTATAC TCCTACTTGG CACTGCAAAC CTGCTCGCAG
25 2451 CACCAGCCGG TGGACTTGCC ATCCAGCTCT CAGCTTCCAC TGCTCCCTT
2501 GTTCCCGGCC GGCTGGCTGC CTCCCCGTGC TGTGTCCAGC ACGGCCAACA
2551 ACGTCAGACC CTCAGAGACG CCCAAGGGGC TTCCAGAGGT GGCCGCTTCT
2601 CTATTTTTTC CTGATTGTGG CTGAGAGAGA TGATTACTGC TTTGACACTT
2651 CCTTTCTCTA AAAGAAAAAT AGTTTGATAG TATATTTTGA ATATAGATGC
30 2701 TCTTATAGTC AGATTGGGAA TTGAACCTGA ATATTGGGTC ATATGTTTGT
2751 GTTGTTGCTG TAGTCTATCA TGACTTTTTT CTTTCTGCAT TTTCTTAAA
2801 AAAAAAAAAA AGATGGCCTT CAAAAGTGTG TTCTCAATGT TGTATGAACC
2851 TCCTTCACAT GAGTTCGGTT GTTGTCTCTC TTCAAAGACT CTTCAACCCA
2901 CAAAGAAGCA ACTAAATGTT TCTCTAAGTT TAATTTTCTA GCGTGTGTT
35 2951 GTCTTACCTT TTTAACCTTA CCATAATATT TCTGTTAAC TTTACATTTA
3001 ATATACCAAT GTGTGTAAGT ATACAGAGAA AAATCTGTTT GTAAAGTAAA
3051 ATTTATATAT AATATATGTA ATCAAAGATA CATATGTTAT ATATACATAT
3101 GTGGATGTAT GACTTATTTT TCCTTATCCA CAGATTTCAG CTACCATGTA
3151 TATATAAATA AACTTATTTT ATTAGCCAGA GAAAAAAA AAAAAAAA
40

```

BLAST Results

45 No BLAST result

Medline entries

50

No Medline entry

55

Peptide information for frame 2

ORF from 29 bp to 1390 bp; peptide length: 454

Category: putative protein

Classification: Transmembrane proteins unclassified

```

5      1 MATTSSKPEG RPRGQAAPTI LLTKPPGATS RPTTAPPRTT TRRPPRPPGS
      51 SRKGAGNSSR PVPPAPGGHS RSKEGQRGRN PSSTPLGQKR PLGKIFQIYK
     101 GNFTGSVEPE PSTLTPRTPL WGYSSSPQPQ TVAATTVPSN TSWAPTTTSL
     151 GPAKDKPGLR RAAQGGGSTF TSQGGTPDAT AASGAPVSPQ AAPVPSQRPH
     201 HGDPAQDGP SH SDSWLTVTPG TSRPLSTSSG VFTAATGPTP AAFDTSVSAP
     251 SQGIPQGA ST TPQAPTHPSR VSESTISGAK EETVATLTMT DRVPSPLSTV
    10  301 VSTATGNFLN RLVPA GTWKP GTAGNISHVA EGDKPQHRAT ICLSKMDIAW
     351 VILAISVPIS SCSVLLTVCC MKRKKKTANP ENNLSYWNNT ITMDYFN RHA
     401 VELPREIQSL ETS EDQLSEP RSPANGDYRD TGMVLVNPFC QETLFVGN DQ
     451 VSEI

```

15

BLASTP hits

No BLASTP hits available

20

Alert BLASTP hits for DKFZphamy2_24c8, frame 2

No Alert BLASTP hits found

25

Pedant information for DKFZphamy2_24c8, frame 2

Report for DKFZphamy2_24c8.2

30

```

[LENGTH] 463
[MW]      48277.84
[pI]      9.80
[FUNCAT]  78 classification not yet clear-cut [S. cerevisiae,
35  YJR151c] 2e-04
[BLOCKS]  PRO0912F
[BLOCKS]  BP03696F
[KW]      TRANSMEMBRANE 1
[KW]      LOW_COMPLEXITY 15.55 %

```

40

```

SEQ LSTGPAPAAMATTSSKPEG RPRGQAAPTILLTKPPGATS RPTTAPPRTT TRRPPRPPGSS
SEG .....
PRD cccccchhhhhhhhhcccccccccccccccccccccccccccccccccccccccccccc
45 MEM .....

```

```

SEQ RKGAGNSSR PVPPAPGGHS RSKEGQRGRN PSSTPLGQKR PLGKIFQIYK GNFTGSVEPEP
SEG x.....
PRD ccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
50 MEM .....

```

```

SEQ STLTPRTPLWGYSSSPQPQ TVAATTVPSN TSWAPTTTSLGPAKDKPGLR RAAQGGGSTFT
SEG .....
PRD ccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
55 MEM .....

```

```

SEQ SQGGTPDATAASGAPVSPQAAPVPSQRPH HGDPAQDGP SH SDSWLTVTPG TSRPLSTSSGV
SEG xxxxxx.....

```

[illegible][illegible][illegible]

```

20  SEQ  TSEDQLSEPRSPANGDYRDTGMVLVNPFCQETLFGNDQVSEI
    SEG  .....
    PRD  ccccccccccccccccccccccccccccccccccccccccccc
    MEM  .....

```

25 (No Prosite data available for DKFZphamy2_24c8.2)

(No Pfam data available for DKFZphamy2_24c8.2)

DKFZphamy2_24k15

5 group: amygdala derived

DKFZphamy2_24k15 encodes a novel 279 amino acid protein with weak similarity to pecanex of *Drosophila melanogaster*.

10 Pecanex is a maternal-effect neurogenic gene, involved in differentiation processes in the developing central nervous system. DKFZphamy2_24k15.p3 seems to be expressed ubiquitously.

15 The new protein can find application in studying the expression profile of amygdala-specific genes and as a new marker for amygdala cells.

similarity to pecanex (*Drosophila melanogaster*)

20 probably complete cds.

Sequenced by GBF

25 Locus: unknown

Insert length: 1464 bp

Poly A stretch at pos. 1445, polyadenylation signal at pos. 1421

30

```

  1 AAGGAAAACA AGAGGACATG CCATATATTC CTCTCATGGA GTTCAGTTGT
  51 TCACATTCTC ACTTAGTATG CTTACCCGCA GAGTGGAGGA CTAGCTGTAT
 101 GCCCAGTTCC AAAATGAAGG AGATGAGCTC GTTATTTCCA GAAGACTGGT
 151 ACCAATTTGT TCTAAGGCAG TTGGAATGTT ATCATT CAGA AGAGAAGGCC
 35 201 TCAAATGTAC TGGAAGAAAT TGCCAAGGAC AAAGTTTTAA AAGACTTTTA
 251 TGTTTCATACA GTAATGACTT GTTATTTTAG TTTATTTGGA ATAGACAATA
 301 TGGCTCCTAG TCCTGGTCAT ATATTGAGAG TTTACGGTGG TGTTTTGCCT
 351 TGGTCTGTTG CTTTGGACTG GCTCACAGAA AAGCCAGAAC TGTTTCAACT
 401 AGCACTGAAA GCATT CAGGT ATACTCTGAA ACTAATGATT GATAAAGCAA
 40 451 GTTTAGGTCC AATAGAAGAC TTTAGAGAAC TGATTAAGTA CCTTGAAGAA
 501 TATGAACGTG ACTGGTACAT TGGTTTTGGTA TCTGATGAAA AGTGGAAAGGA
 551 AGCAATTTTA CAAGAAAAGC CATACTTGTT TTCTCTGGGG TATGATTCTA
 601 ATATGGGAAT TTACACTGGG AGAGTGCTTA GCCTTCAAGA ATTATTGATC
 651 CAAGTGGGAA AGTTAAATCC TGAAGCTGTT AGAGGTCAGT GGGCCAATCT
 45 701 TTCATGGGAA TTACTTTATG CCACAAACGA TGATGAAGAA CGTTATAGTA
 751 TACAAGCTCA TCCACTACTT TTAAGAAATC TTACGGTACA AGCAGCAGAA
 801 CCTCCCCTGG GATATCCGAT TTATTCTTCA AAACCTCTCC ACATACATTT
 851 GTATTAGAGC TCATTTTGAC TGTAATGTCA TCAAATGCAA TGTTTTTATT
 901 TTTTCATCCT AAAAAAGTAA CTGTGATTCT TGTAACCTGA GGACTTCTCC
 50 951 ACACCCCCAT TCAGATGCCT GAGAACAGCT AAGCTCCGTA AAGTTGGTTC
 1001 TCTTAGCCAT CTTAATGGTT CTA AAAAACA GCAAAAACAT CTTTATGTCT
 1051 AAGATAAAAG AACTATTTGG CCAATATTTG TGCCCTCTGG ACTTTAGTAG
 1101 GCTTTGGTAA ATGTGAGAAA ACTTTTG TAG AATTATCATA TAATGAATTT
 1151 TGTAATGCTT TCTTAAATGT GTTATAGGTG AATTGCCATA CAAAGTTAAC
 55 1201 AGCTATGTAA TTTTACATA CTTAAGAGAT AAACATATCA GTGTTCTAAG
 1251 TAGTGATAAT GGATCCTGTT GAAGGTTAAC ATAATGTGTA TATATTTGTT
 1301 TGAAATATAA TTTATAGTAT TTTCAAATGT GCTGATTTAT TTTGACATCT
 1351 AATATCTGAA TGTTTTTGTA TCAAGTAGTT TGTTTTCATA GACTTCAATT

```

1401 CATAAACTTT AAAAAACTTT TAATAAAATA TTTTCCTTCC TTTTCAAAAA
 1451 AAAAAAAAAA AAAA

5

BLAST Results

Entry AC007939 from database EMBLNEW:

Homo sapiens clone 422_H_5, WORKING DRAFT SEQUENCE, 5 unordered pieces.

10

Score = 4116, P = 0.0e+00, identities = 840/858

3 exons

15

Medline entries

No Medline entry

20

Peptide information for frame 3

25

ORF from 18 bp to 854 bp; peptide length: 279

Category: similarity to known protein

Classification: unclassified

30

1 MPYIPLMEFS CSHSHLVCLP AEWRTSCMPS SKMKEMSSLF PEDWYQFVLR
 51 QLECYHSEEK ASNVLEEIAK DKVLKDFYVH TVMTCYFSLF GIDNMAPSPG
 101 HILRVYGGVL PWSVALDWLT EKPELFQLAL KAFRYTLKLM IDKASLGPIE
 151 DFRELIKYLE EYERDWYIGL VSDEKWKEAI LQEKPYLFSL GYDSNMGIYT
 201 GRVLSLQELL IQVGKLNPEA VRGQWANLSW ELLYATNDDE ERYSIQAHPL
 35 251 LLRNLTVQAA EPPLGYPIYS SKPLHIHLY

BLASTP hits

40

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_24k15, frame 3

45 No Alert BLASTP hits found

Pedant information for DKFZphamy2_24k15, frame 3

50

Report for DKFZphamy2_24k15.3

[[LENGTH]] 284

[[MW]] 33066.31

55

[[pI]] 5.17

[[HOMOL]] TREMBL:AF067608_11 gene: "B0511.12";

Caenorhabditis elegans cosmid B0511. 2e-13

[[KW]] Alpha_Beta

(No Pfam data available for DKFZphamy2_24k15.3)

DKFZphamy2_2a13

5 group: amygdala derived

DKFZphamy2_2a13 encodes a novel 440 amino acid protein without similarity to known proteins.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motif.

The new protein can find application in studying the expression profile of amygdala-specific genes.

15

putative protein

perhaps complete cds.

20

Sequenced by MediGenomix

Locus: /map="1bpl3.3"

25 Insert length: 2584 bp

Poly A stretch at pos. 2562, polyadenylation signal at pos. 2545

```

30      1 GTTCCTGAGG ACGTGCTACG GGGGCAGCTT CCTGGTACAC GAGTCGTTCC
      51 TCTACAAGCG GGAGAAGGCT GTCGGGGACA AGGTGTATTG GACCTGCCGG
     101 GACCACGCGC TGCACGGCTG CCGGAGCCGG GCCATCACCC AGGGACAGCG
     151 GGTGACTGTG ATGCGTGGGC ACTGCCACCA GCCCGATATG GAGGGCCTGG
     201 AAGCCCGGCG GCAGCAGGAG AAGGCCGTGG AGACGCTGCA GGCTGGGCAG
     251 GACGGCCCTG GGAGCCAAGT GGACACGCTG CTCCGAGGCG TGGATAGTTT
     35  301 GCTCTACCGC AGGGGTCCGG GTCCCTGAC TCTCACCAGG CCTCGGCCCCA
     351 GAAAGCGAGC AAAGGTCGAA GACCAGGAGC TGCCAACCCA GCCCGAGGCC
     401 CCAGACGAGC ACCAGGACAT GGACGCAGAC CCGGGAGGCC CTGAGTTCCT
     451 GAAGACGCCC CTGGGGGGCA GCTTCCTGGT GTACGAGTCC TTCCTCTACC
     501 GCGGGGAGAA GCGGCTGGG GAGAAGGTGT ATTGGACCTG CCGGGACCAG
     40  551 GCGCGCATGG GCTGCCGCAG CCGCGCCATC ACCCAGGGCC GACGGGTGAC
     601 TGTCATGCGT GGTCACTGCC ACCCGCCCGA CCTGGGAGGC CTGGAGGCC
     651 TGAGGCAGCG GGAGAAACGC CCAACACGG CGCAGCGGGG GAGCCAGGC
     701 GCTGGCCTCT CTTTCCAGTG GCTCTTCCGG ATCCTGCAGC TTTTGGGTCA
     751 TGCTCCTGTG CTGCTGTGCC CCTCAGGGTC CTCCTGCCTC CCGAGCCTCC
     45  801 CTGCTCCACA TGGCCCCCTG CAGCCCTCT CCATCCCTCT TGAAGGAGGC
     851 CCCGAGTTCC TGAAGACGCC CCTGGGGGGC AGCTTCCTGG TGTACGAGTC
     901 CTTCTCTAC CGGCGGGAGA AGGCGGCCGG GGAGAAGGTG TATTGGACCT
     951 GCGGGGACCA GGCCCGCATG GGCTGCCGCA GCGCGCCAT CACCCAGGGC
    1001 CGGCGGGTCA TGGTCATGCG CAGGCACTGC CACCCACCGG ACCTGGGCGG
    50 1051 CCTGGAGGCC CTGCGGCAGC GGGAGCACTT CCCCACCTG GCGCAGTGGG
    1101 ACAGCCAGAG TCCTCTCCGG CCCCTGGAGT TCCTGAGGAC TTCCCTGGGG
    1151 GGCAGGTTCC TGGTGACGA GTCCTTCCTC TACAGGAAGG AGAAGGCGGC
    1201 TGGGGAGAA GGTACTGGA TGTGCCGGGA CCAGGCTCGG CTGGGCTGCC
    1251 GCAGCCGCGC CATAACCCAG GGCCACCGCA TCATGGTCAT GCGCAGCCAC
    55 1301 TGCCATCAGC CTGACCTGGG AGGCTTGGAG GCCTTGAGGC AACGGGAGCG
    1351 GCTCCCCACC ACGGCCAGC AGGAGGACCC AGAAAAGATT CAAGTTCAGC
    1401 TGTGCTTCAA GACGTGTTCT CCTGAAAGCC AGCAGATTTA TGGGGACATC
    1451 AAAGACGTCA GACTGGATGG CGAGTCCCAG TGAGGCGATG TGGGCAGAGG

```

```

1501 AGCTCCGAGC CGCCACCCA AGGTGGCTTC ACATCCACAC AGGCACTTCC
1551 CATCCACCTA GGTTTGGCTT AGCAGAAACT TCTTTTCATT CTTCCAAAGC
1601 ATCGATGGTC TTCGCGTCTC CTCAGGAGGT CTCCAGGAG GAATTCCTGG
1651 ATGGTGTCTT CATGTCGGCG GAGAACAGTG CTCAGAGCTG GCGCTTGCAg
5 1701 ACGCAGCTGT CGTGGGGCAG GCGGGTGGCG CCTTCCTGAC CTTTGGAAGA
1751 CATGACAAAG CTGCCTGGAC ACGGACGCC CTGCTGTACG GCCACAGCAC
1801 CCCTGGGTTT GCAGAGCACG CAGCCTTCTT AGGGCTTTCC ACCTGGCGAG
1851 GCCCCGCTCT GCTCAGCACG GTGCAAAGTG AATGCTGCTG TCTTGGAGCC
1901 TGGGCACGTT TGGGGAAGTT CCTGCTTCAA ACTGAGCTGC CCCGCATAGG
10 1951 CCAGGTCAAC CCACACCAAT CTTTCTGGA CAGGTGCTGG GTAGGCCTTC
2001 CTGGTCTCTG GCCGCTGCT GCCAGGGTGT GGCCATCCCC AGCAACCGGA
2051 GCCGGCCAAA CCAGAGGCCT GCCTCCGCAC TCCACACTTT CTTTCTGTG
2101 CTCCTTCCAA GTTAAATTAA ACCCCTCTC CACGATTCCC ACGGCAGGCG
2151 TCATTCCCGA GATGGGAGCC AGTCCAGGGG TCAGCAGGAG CCAGCGCTGG
15 2201 GCACACGTGC CCTGGCTGAG GCCAGCGGCA TCCTGGGTGG CCCAGGTCCA
2251 TCCTGGGCAG CAAAGGCGTG TCCCCTTCTG TCAGACAGCT TCACAGAGTG
2301 TGGCTTCACC AGTCAGAGGG AGCAGTCCGG AGAGGCAAGA TGACCCACC
2351 GGGACTGCAG AGCCTCCTCC TTAATAACAA GGACCTGTCC GCAGCCGCGA
2401 GGTCTTTCAC TCCCACCCTG TAATTGTGGG GGGAGTGCCA GCAACAGGCC
20 2451 TGTCCCCTGG CAAGTTGGCC ACGGAACCCA CCATGCACTG CAAGGCTGTG
2501 ACAGCCTGGG CACCCCTGCT TCTCCTCTGC TTGTACGGT CCCCCAATAA
2551 ATCCTATTTT CCATCAAAAA AAAAAAAAAA AAAA

```

25 BLAST Results

No BLAST result

30 Medline entries

No Medline entry

35 Peptide information for frame 2

40 ORF from 161 bp to 1480 bp; peptide length: 440
Category: putative protein
Classification: no clue

```

45 1 MRGHCHQPD M EGLEARRQAE KAVETLQAGQ DGPGSQVDTL LRGVDSL LLYR
51 RGPGLTLTR PRPRKRAKVE DQELPTQPEA PDEHQDMDAD PGGPEFLKTP
101 LGG SFLVYES FLYRREKAAG EK VYWT CRDQ ARMGCRSRAI TQGRRTVMR
151 GHCHPPDLGG LEALRQREKR PNTAQRGSPG AGLSFQWLFR ILQLLGHAPV
201 LLCPSGSSCL PSLPAPHGPC PALSIPLEGG PEFLKTPLGG SFLVYESFLY
50 251 RREKAAGEKV YWTCRDQARM GCRSRAITQG RRVMMRRHC HPPDLGGLEA
301 LRQREHFPNL AQWDSPDPLR PLEFLRTSLG GRFLVHESFL YRKEKAAGEK
351 VYWMCRDQAR LGCRSRAITQ GHRIMVMRSH CHQPDLAGLE ALRQERLPT
401 TAQQEDPEKI QVQLCFKTC S PESQRIYGDI KDVRLDGESQ

```

55 BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_2a13, frame 2

5 No Alert BLASTP hits found

Pedant information for DKFZphamy2_2a13, frame 2

10 Report for DKFZphamy2_2a13.2

15 [LENGTH] 493
[MW] 55840.13
[pI] 9.33
[KW] Alpha_Beta
[KW] LOW_COMPLEXITY 6.29 %

20 SEQ FLRTCYGGSFLVHESFLYKREKAVGDKVYWTCDHALHGCRSRAITQGQRTVVMRGHCHQ
SEG
PRD cccccccccceecchhhhhhhhhccceeeeeeccccccccccccceeeecceeeeeecccccc

25 SEQ PDMEGLEARRQDEKAVETLQAGQDGPQSQVDTLLRGVDSLLYRRGPGPLTLTRPRPRKRA
SEGxxxxxxxxxxxxxxxxxxxxx...
PRD cccchhhhhhhhhhhhhhhhhhhccccccccccccccccccccceeeecceeeecceccccchhh

30 SEQ KVEDQELPTQPEAPDEHQDMDADPGGPEFLKTPLGGSFLVYESFLYRREKAAGEKVYWTC
SEG
PRD hhhhhccccccccccccccccccccccccccccccccccccceeehhhhhhhhhhhhccceeeec

35 SEQ RDQARMGCRSRAITQGRRTVVMRGHCHPPDLGGLEALRQREKRPNTAQRGSPGAGLSFQW
SEG
PRD cchhhhhccceeeccccceeeeeeccccccccccchhhhhhhhhccccccccccccccchhhhh

40 SEQ LFRILQLLGHAPVLLCPSGSSCLPSLPAPHGPCPALSIPLEGGPEFLKTPLGGSFLVYES
SEGxxxxxxxxxxxxxxxxxxxxx...
PRD hhhhhhhhhccceeeccceeehh

45 SEQ FLYRREKAAGEKVYWTCRDQARMGCRSRAITQGRVMVMRRHCHPPDLGGLEALRQREHF
SEG
PRD hhhhhhhhhccceeeecchhhhhccceeeccccceeeeeeccccccccccchhhhhhhhhc

50 SEQ PNLAQWDSPDPLRPLEFLRTSLGGRFLVHESFLYRKEKAAGEKVYWMCRDQARLGCRSRA
SEG
PRD cccccccccccchhhhhhhhhccceeeecchhhhhhhhhccceeeecchhhhhhhhhcccc

55 SEQ ITQGHRIMVMRSHCHQPDLAGLEALRQRERLPTTAQQEDPEKIQVQLCFKTCSPESQDIY
SEG
PRD cccccceeeeeeccccccccccchhhhhhhhhhhhhhhccccccccceeehhhhcccccccccc

SEQ GDIKDVRLDGESQ
SEG
PRD ccccccccccccc

(No Prosite data available for DKFZphamy2_2a13.2)

(No Pfam data available for DKFZphamy2_2a13.2)

DKFZphamy2_2b19

5 group: differentiation/development

DKFZphamy2_2b19 encodes a novel 789 amino acid protein which originates

from TXBP151 mRNA by alternative splicing.

10

It is ubiquitously expressed. The mRNA is also subject to alternative polyadenylation. Overexpression of TXBP151 in NIH3T3 cells causes inhibition of

15

apoptosis induced by tumour necrosis factor (TNF). It binds to A20, which is also an inhibitor of cell death by a yet unknown mechanism.

The new protein can find application in modifying/blocking apoptotic pathways and therefore serve as a tool in diagnosis of cancer predisposition and as a tool in cell culture.

20

TXBP151, differentially spliced

25

differential splicing
differential polyadenylation

Sequenced by MediGenomix

30

Locus: /map="7p15"

Insert length: 3028 bp

Poly A stretch at pos. 2885, polyadenylation signal at pos. 2868

35

	1	GAAGAGGTTT	GGCGGCTGAT	GGCGGATCAG	GATCGGAAGC	CTGCGTAACT
	51	TTCTCCCTTG	ATCCGGGAGT	CTTTCCACTG	GATTCACAAT	GACATCCTTT
	101	CAAGAAGTCC	CATTGCAGAC	TTCCAACCTT	GCCCATGTCA	TCTTTCAAAA
	151	TGTGGCCAAG	AGTTACCTTC	CTAATGCACA	CCTGGAATGT	CATTACACCT
40	201	TAACTCCATA	TATTCATCCA	CATCCAAAAG	ATTGGGTTGG	TATATTCAAG
	251	GTTGGATGGA	GTACTGCTCG	TGATTATTAC	ACGTTTTTAT	GGTCCCCTAT
	301	GCCTGAACAT	TATGTGGAAG	GATCAACAGT	CAATTGTGTA	CTAGCATTCC
	351	AAGGATATTA	CCTTCCAAAT	GATGATGGAG	AATTTTATCA	GTTCTGTTAC
	401	GTTACCCATA	AGGGTGAAAT	TCGTGGAGCA	AGTACACCTT	TCCAGTTTCG
45	451	AGCTTCTTCT	CCAGTTGAAG	AGCTGCTTAC	TATGGAAGAT	GAAGGAAATT
	501	CTGACATGTT	AGTGGTGACC	ACAAAAGCAG	GCCTTCTTGA	GTTGAAAATT
	551	GAGAAAACCA	TGAAAAGAAA	AGAAGAAGTG	TTAAAGTTAA	TTGCCGTTCT
	601	GGAAAAGAG	ACAGCACAAC	TTGAGAACAA	AGTTGGGAGA	ATGGAAAGAG
	651	AACTTAACCA	TGAGAAAGAA	AGATGTGACC	AACTGCAAGC	AGAACAAAAG
50	701	GGTCTTACTG	AAGTAACACA	AAGCTTAAAA	ATGGAAAATG	AAGAGTTTAA
	751	GAAGAGGTTT	AGTGATGCTA	CATCCAAAGC	CCATCAGCTT	GAGGAAGATA
	801	TTGTGTCACT	AACACATAAA	GCAATTGAAA	AAGAAACCGA	ATTAGACAGT
	851	TTAAAGGACA	AACTCAAGAA	GGCACAACAT	GAAAGAGAAC	AACTTGAATG
	901	TCAGTTGAAG	ACAGAGAAGG	ATGAAAAGGA	ACTTTATAAG	GTACATTTGA
55	951	AGAATACAGA	AATAGAAAAT	ACCAAGCTTA	TGTCAGAGGT	CCAGACTTTA
	1001	AAAAATTTAG	ATGGGAACAA	AGAAAGCGTG	ATTACTCATT	TCAAAGAAGA
	1051	GATTGGCAGG	CTGCAGTTAT	GTTTGGCTGA	AAAGGAAAAT	CTGCAAAGAA
	1101	CTTTCCTGCT	TACAACCTCA	AGTAAAGAAG	ATACTTGTTT	TTTAAAGGAG

1151 CAACTTCGTA AAGCAGAGGA ACAGGTTTCAG GCAACTCGGC AAGAAGTTGT
1201 CTTTCTGGCT AAAGAACTCA GTGATGCTGT CAACGTACGA GACAGAACGA
1251 TGGCAGACCT GCATACTGCA CGCTTGGAAG ACGAGAAAGT GAAAAAGCAG
1301 TTAGCTGATG CAGTGGCAGA ACTTAAACTA AATGCTATGA AAAAAAGATCA
5 1351 GGACAAGACT GATACACTGG AACACGAACT AAGAAGAGAA GTTGAAGATC
1401 TGAAACTCCG TCTTCAGATG GCTGCAGACC ATTATAAAGA AAAATTTAAG
1451 GAATGCCAAA GGCTCCAAAA ACAAAATAAAC AAACTTTCAG ATCAATCAGC
1501 TAATAATAAT AATGTCTTCA CAAAGAAAAC GGGGAATCAG CAGAAAGTGA
1551 ATGATGCTTC AGTAAACACA GACCCAGCCA CTTCTGCCTC TACTGTAGAT
10 1601 GTAAAGCCAT CACCTTCTGC AGCAGAGGCA GATTTTGACA TAGTAACAAA
1651 GGGGCAAGTC TGTGAAATGA CCAAGAAAT TGCTGACAAA ACAGAAAAGT
1701 ATAATAAATG TAAACAATG TTGCAGGATG AGAAAGCAAA ATGCAATAAA
1751 TATGCTGATG AACTTGCAAA AATGGAGCTG AAATGGAAAG AACAAGTGAA
1801 AATTGCTGAA AATGTAAAAC TTGAACTAGC TGAAGTACAG GACAATTATA
15 1851 AAGAACTTAA AAGGAGTCTA GAAAATCCAG CAGAAAGGAA AATGGAAGGT
1901 CAGAATTCCC AGAGTCCTCA ATGTTTCAAA ACATGCTCAG AGCAAAATGG
1951 TTATGTTCTC ACATTGTCAA ATGCACAACC AGTTCTGCAA TATGGTAATC
2001 CTTATGCATC TCAGGAAACA AGAGATGGAG CAGATGGTGC TTTTACCCA
2051 GATGAAATAC AAAGGCCACC TGTGAGAGTC CCTCTTGGG GACTGGAAGA
20 2101 CAATGTTGTC TGCAGCCAGC CTGCTCGAAA CTTTAGTCGG CCTGATGGCT
2151 TAGAGGACTC TGAGGATAGC AAAGAAGATG AGAATGTGCC TACTGCTCCT
2201 GATCCTCCAA GTCAACATTT ACGTGGGCAT GGGACAGGCT TTTGCTTTGA
2251 TTCCAGCTTT GATGTTTACA AGAAGTGTCC CCTCTGTGAG TTAATGTTTC
2301 CTCCTAACTA TGATCAGAGC AAATTTGAAG AACATGTTGA AAGTCACTGG
25 2351 AAGGTGTGCC CGATGTGCAG CGAGCAGTTC CCTCCTGACT ATGACCAGCA
2401 GGTGTTTGAA AGGCATGTGC AGACCCATTT TGATCAGAAT GTTCTAAATT
2451 TTGACTAGTT ACTTTTTATT ATGAGTTAAT ATAGTTTAGC AGTAAAAAAA
2501 AAAAAAAAAA ACCACACCTA AAATAGACCA CTGAGGAGAC CATAGAGCGG
2551 ATGCTTTTCAT GCACCTTTTA CTGCACTTTC TGACCAGGAG CTACTTTGAG
30 2601 TTTGGTGTTA CTAGGATCAG GGTGAGTCTT TGGCTTATCA ATAAATTTTA
2651 ATCTCTGTTA ATCTTACCTG CTTTAAAAAA AAGTTCTTGT GTGTTTCGTAT
2701 CTTTATTTAT TCCCTAGTTT GCAGAACTGT CTGAATAAAG GATACAAGGA
2751 TTATTTCAAT GTTACTGCAC TGAAAAACGT GTATGTATTA GTGTGCTAGA
2801 TTATTTAGCA GAATATTAC AAGTTTCTGT TGACCTTGTT GATTGAGCAT
35 2851 GACTACTAAA TATTATGTAA TAAAAAGCAT TTGTCATAAC AAAAAAAAAA
2901 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
2951 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
3001 AAAAAAAAAA AAAAAAAAAA AAAAAAAA

40

BLAST Results

No BLAST result

45

Medline entries

50

99361984:

De Valck D, Jin DY, Heyninck K, Van de Craen M, Contreras R,
Fiers W,Jeang KT, Beyaert R.; The zinc finger protein A20 interacts with
a

55

novel

anti-apoptotic protein which is cleaved by specific caspases.
Oncogene

1999 Jul 22;18(29):4182-90

5 Peptide information for frame 2

ORF from 89 bp to 2455 bp; peptide length: 789

Category: known protein

10 Classification: Cell division

```

1   1 MTSFQEVPLQ TSNFAHVIFQ NVAKSYLPNA HLECHYTLTP YIHPHPKDQWV
5   51 GIFKVGWSTA RDYYTFLWSP MPEHYVEGST VNCVLAFAQGY YLPNDGGEFY
10  101 QFCYVTHKGE IRGASTPFQF RASSPVEELL TMEDEGNSDM LVVTTKAGLL
15  151 ELKIEKTMKE KEELLKLIIV LEKETAQLRE QVGRMERELN HEKERCDQLQ
20  201 AEQKGLTEVT QSLKMENEEF KKRFSDATSK AHQLEEDIVS VTHKAIEKET
25  251 ELDSLKDCLK KAQHEREQLE CQLKTEKDEK ELYKVHLKNT EIENTKLMSE
30  301 VQTLKLNLDGN KESVITHFKE EIGRLQLCLA EKENLQRTFL LTTSSKEDTC
35  351 FLKEQLRKAE EQVQATRQEV VFLAKELSDA VNVDRDTMAD LHTARLENEK
40  401 VKKQLADAVA ELKLNAMKKD QDKTDTLEHE LRREVEDLKL RLQMAADHYK
45  451 EKFKECQRLQ KQINKLSDAQ ANNNNVFTKK TGNQQKVNDQ SVNTDPATSA
50  501 STVDVKPSPS AAADFDIVT KGQVCEMTKE IADKTEKYNK CKQLLQDEKA
55  551 KCNKYADELA KMELKWKEQV KIAENVKLEL AEVQDNYKEL KRSLENPAER
60  601 KMEGQNSQSP QCFKTCSEQN GYVLTLSNAQ PVLQYGNPYA SQETRDGADG
25  651 AFYPDEIQRP PVRVPSWGLE DNVVCSQPAR NFSRPDGLD SEDSKEDENV
70  701 PTAPDPPSQH LRGHGTGFCF DSSFVHKKC PLCELMFPPN YDQSKFEEHV
75  751 ESHWKVCPMC SEQFPPDYDQ QVFERHVQTH FDQNVLNFD

```

30

BLASTP hits

No BLASTP hits available

35 Alert BLASTP hits for DKFZphamy2_2b19, frame 2

TREMBL:HS338211_1 product: "tax1-binding protein TXBP151"; Homo sapiens tax1-binding protein TXBP151 mRNA, complete cds.; N = 2, Score

40 = 2948, P = 0

>TREMBL:HS338211_1 product: "tax1-binding protein TXBP151"; Homo sapiens

45 tax1-binding protein TXBP151 mRNA, complete cds.
Length = 747

HSPs:

50 Score = 2948 (442.3 bits), Expect = 0.0e+00, Sum P(2) = 0.0e+00
Identities = 575/603 (95%), Positives = 576/603 (95%)

Query: 1

MTSFQEVPLQTSNFAHVIFQNVAKSYLPNAHLECHYTLTPYIHPHPKDQWVGIFKVGWSTA 60

55

MTSFQEVPLQTSNFAHVIFQNVAKSYLPNAHLECHYTLTPYIHPHPKDQWVGIFKVGWSTA

Sbjct: 1

MTSFQEVPLQTSNFAHVIFQNVAKSYLPNAHLECHYTLTPYIHPHPKDQWVGIFKVGWSTA 60

Query: 61
 RDYYTFLWSPMPEHYVEGSTVNCVLAFQGYLPLNDDGEFYQFCYVTHKGEIRGASTPFQF 120
 5 RDYYTFLWSPMPEHYVEGSTVNCVLAFQGYLPLNDDGEFYQFCYVTHKGEIRGASTPFQF
 Sbjct: 61
 RDYYTFLWSPMPEHYVEGSTVNCVLAFQGYLPLNDDGEFYQFCYVTHKGEIRGASTPFQF 120

Query: 121
 10 RASSPVEELLTMEDEGNSDMLVVTTKAGXXXXXXXXXXXXXXXXXXXXXXXXXXXXTAQLRE 180
 RASSPVEELLTMEDEGNSDMLVVTTKAG
 TAQLRE
 Sbjct: 121
 RASSPVEELLTMEDEGNSDMLVVTTKAGLLELLEKIEKTMKEKEELLKLIIVLEKETAQLRE 180
 15

Query: 181
 QVGRMERELNHEKERCDQLQAEQKGLTEVTQSLKMENEEFKKRFSDATSKAHQLEEDIVS 240
 QVGRMERELNHEKERCDQLQAEQKGLTEVTQSLKMENEEFKKRFSDATSKAH
 +EEDIVS
 20 Sbjct: 181
 QVGRMERELNHEKERCDQLQAEQKGLTEVTQSLKMENEEFKKRFSDATSKAHHVEEDIVS 240

Query: 241
 25 VTHKAIEKETELDSLKDKLKAQHEREQLEQKTEKDEKELYKVHLKNTIEIENTKLMSE 300
 VTHKAIEKETELDSLKDKLKAQHEREQLEQKTEKDEKELYKVHLKNTIEIENTKLMSE
 Sbjct: 241
 VTHKAIEKETELDSLKDKLKAQHEREQLEQKTEKDEKELYKVHLKNTIEIENTKLMSE 300

Query: 301
 30 VQTLKNLDGNKESVITHFKEEIGRLQLCLAEKENLQRTFLLTTSSKEDTCFLKEQLRKAE 360
 VQTLKNLDGNKESVITHFKEEIGRLQLCLAEKENLQRTFLLTTSSKEDTCFLKEQLRKAE
 Sbjct: 301
 35 VQTLKNLDGNKESVITHFKEEIGRLQLCLAEKENLQRTFLLTTSSKEDTCFLKEQLRKAE 360

Query: 361
 EQVQATRQEVVFLAKELSDAVNVRDRTMADLHTARLENEKVKKQLADAVAELKLNAMKKD 420
 40 EQVQATRQEVVFLAKELSDAVNVRDRTMADLHTARLENEKVKKQLADAVAELKLNAMKKD
 Sbjct: 361
 EQVQATRQEVVFLAKELSDAVNVRDRTMADLHTARLENEKVKKQLADAVAELKLNAMKKD 420

Query: 421
 45 QDKTDTLEHELRRVEDLKLRLQMAADHYKEKFKECQRLQKQINKLSDQSANNNNVFTKK 480
 QDKTDTLEHELRRVEDLKLRLQMAADHYKEKFKECQRLQKQINKLSDQSANNNNVFTKK
 Sbjct: 421
 QDKTDTLEHELRRVEDLKLRLQMAADHYKEKFKECQRLQKQINKLSDQSANNNNVFTKK 480
 50

Query: 481
 TGNQKQVNDASVNTDPATSASTVDVKPSPSAAEADFIVTKGQVCEMTKEIADKTEKYNK 540
 TGNQKQVNDASVNTDPATSASTVDVKPSPSAAEADFIVTKGQVCEMTKEIADKTEKYNK
 55 Sbjct: 481
 TGNQKQVNDASVNTDPATSASTVDVKPSPSAAEADFIVTKGQVCEMTKEIADKTEKYNK 540

Query: 541
CKQLLQDEKAKCNKYADELAKMELKWKEQVKIAENVKLELAEVQDNKELKRSLENPAER 600

5 Sbjct: 541
CKQLLQDEKAKCNKYADELAKMELKWKEQVKIAENVKLELAEVQDNKELKRSLENPAER 600

Query: 601 KME 603
KME

10 Sbjct: 601 KME 603

Score = 831 (124.7 bits), Expect = 0.0e+00, Sum P(2) = 0.0e+00
Identities = 147/153 (96%), Positives = 149/153 (97%)

15 Query: 637
NPYASQETRDGADGAFYPDEIQRPVVRVPSWGLEDNVVCSPARNFSRPDGLEDSSEDSKE 696
NP A ++

DGADGAFYPDEIQRPVVRVPSWGLEDNVVCSPARNFSRPDGLEDSSEDSKE

Sbjct: 596 NP-

20 AERKMEDGADGAFYPDEIQRPVVRVPSWGLEDNVVCSPARNFSRPDGLEDSSEDSKE 654

Query: 697
DENVPTAPDPPSQHLRGHGTGFCFDSSFDVHKKCPLCELMFPPNYDQSKFEEHVESHWKV 756

25 DENVPTAPDPPSQHLRGHGTGFCFDSSFDVHKKCPLCELMFPPNYDQSKFEEHVESHWKV

Sbjct: 655

DENVPTAPDPPSQHLRGHGTGFCFDSSFDVHKKCPLCELMFPPNYDQSKFEEHVESHWKV 714

Query: 757 CPMCSEQFPDPDYDQVFERHVQTHFDQNVLNFD 789

30 CPMCSEQFPDPDYDQVFERHVQTHFDQNVLNFD

Sbjct: 715 CPMCSEQFPDPDYDQVFERHVQTHFDQNVLNFD 747

Score = 104 (15.6 bits), Expect = 9.2e-02, Sum P(2) = 8.8e-02
Identities = 80/351 (22%), Positives = 157/351 (44%)

35 Query: 177 QLR---EQVGRMERELNH-
EKERCDQLQAEQKGLTEVTQSLKMENEEFKRFSDATSKAH 232

QLR EQV +E+ KE D + + + + + + + + + + ENE+ KK+

+DA

40 Sbjct: 355 QLRKAEQVQATRQEVVFLAKELSDAVNVRDRTMADL-
HTARLENEKVKKQLADA----- 408

Query: 233 QLEEDIVSVTHKAIEKETE-
LDSLKDKLKKAQHEREQLECKTEKDEKELYKVHLKNT 291

45 + + A++K+ + D+L+ +L++ E E L+ +L+ D

YK K +

Sbjct: 409 -----VAELKLNAMKKDQDKTDLEHELRR---EVEDLKLRLQMAADH---
YKEKFKECQ 457

50 Query: 292
IENTKLMSEVQTLKNLDGNKESVITHFKEEIGRLQLCLAEKENLQRTFLLTSSKEDTCF 351
+L ++ L + N +V T ++ G Q N T

T++S D

Sbjct: 458 ----RLQKQINKLSDQSANNNNVFT---KKTGNQKQVNDASVN---

55 TDPATSASTVD--- 504

Query: 352 LKEQLRKAEQVQ-
ATRQEVVFLAKELSDAVNVRDRTMADLHTARLENEKVKKQLADAVA 410

+K AE T+ +V + KE++D ++ L + + K
+LA
Sbjct: 505
VKPSPSAAEADFIDIVTKGQVCEMTKEIADKTEKYNKCKQLLQDEKAKCNKYADELAKMEL 564
5
Query: 411 ELKLNAMKKDQDKTDTLE-----HELRRVED-LKLRLQMAAD--
HYKEKFKECQ-RLQK 461
+ K + K + E EL+R +E+ + +++ AD Y ++ +
R+
10 Sbjct: 565
KWKEQVKIAENVKLELAEVQDNYKELKRSLENPAERKMEDGADGAFYPDEIQRPVVRVPS 624
Query: 462 ---QINKLSDQSANNNNVFTKKTG---
NQQKVNDASVNTDPATSASTVDVKPSPSAAEAD 515
15 + N + Q A N F++ G ++ D +V T P + +
+ ++
Sbjct: 625 WGLEDNVVCSPARN---
FSRPDGLEDSKEDENVPTAPDPPSQHLRGHGTGFCFDSS 681
20 Query: 516 FDIVTKGQVCEM 527
FD+ K +CE+
Sbjct: 682 FDVHKKCPLCEL 693

25 Pedant information for DKFZphamy2_2b19, frame 2

Report for DKFZphamy2_2b19.2

30 [LENGTH] 789
[MW] 90877.47
[pI] 5.30
[CHOMOL] TREMBL:HS338211_1 product: "tax1-binding protein
35 TXBP151"; Homo sapiens tax1-binding protein TXBP151 mRNA,
complete cds. 0.0
[FUNCAT] 99 unclassified proteins [S. cerevisiae, YOR216c]
3e-14
[FUNCAT] 08.07 vesicular transport (golgi network, etc.) [S.
40 cerevisiae, YDL058w] 2e-13
[FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae,
YDL058w] 2e-13
[FUNCAT] 09.10 nuclear biogenesis [S. cerevisiae, YDR356w]
4e-13
45 [FUNCAT] 30.04 organization of cytoskeleton [S. cerevisiae,
YDR356w] 4e-13
[FUNCAT] 03.22 cell cycle control and mitosis [S. cerevisiae,
YDR356w] 4e-13
[FUNCAT] 11.04 dna repair (direct repair, base excision repair
50 and nucleotide excision repair) [S. cerevisiae, YKR095w] 7e-12
[FUNCAT] 30.10 nuclear organization [S. cerevisiae, YKR095w]
7e-12
[FUNCAT] 03.25 cytokinesis [S. cerevisiae, YHR023w MY01 -
myosin-1 isoform] 6e-11
55 [FUNCAT] 08.22 cytoskeleton-dependent transport [S. cerevisiae,
YHR023w MY01 - myosin-1 isoform] 6e-11
[FUNCAT] 03.04 budding, cell polarity and filament formation
[S. cerevisiae, YHR023w MY01 - myosin-1 isoform] 6e-11

[FUNCAT] 1 genome replication, transcription, recombination and repair [M. jannaschii, MJ1322] 3e-08
 [FUNCAT] 98 classification not yet clear-cut [S. cerevisiae, YJR134c] 4e-08
 5 [FUNCAT] 03.19 recombination and dna repair [S. cerevisiae, YNL250w] 2e-07
 [FUNCAT] 03.13 meiosis [S. cerevisiae, YNL250w] 2e-07
 [FUNCAT] 03.01 cell growth [S. cerevisiae, YNL079c] 2e-06
 [FUNCAT] 03.07 pheromone response, mating-type determination, sex-specific proteins [S. cerevisiae, YNL079c] 2e-06
 10 [FUNCAT] 08.99 other intracellular-transport activities [S. cerevisiae, YNL079c] 2e-06
 [FUNCAT] 09.13 biogenesis of chromosome structure [S. cerevisiae, YLR086w] 5e-06
 15 [FUNCAT] 11.01 stress response [S. cerevisiae, YPR141c] 2e-05
 [FUNCAT] 06.10 assembly of protein complexes [S. cerevisiae, YPR141c] 2e-05
 [FUNCAT] 03.22.01 cell cycle check point proteins [S. cerevisiae, YGL086w] 2e-05
 20 [FUNCAT] 30.05 organization of centrosome [S. cerevisiae, YPR141c] 2e-05
 [FUNCAT] 08.16 extracellular transport [S. cerevisiae, YOR326w] 1e-04
 [FUNCAT] 09.25 vacuolar and lysosomal biogenesis [S. cerevisiae, YOR326w] 1e-04
 25 [FUNCAT] 30.16 mitochondrial organization [S. cerevisiae, YAL011w] 2e-04
 [FUNCAT] 06.07 protein modification (glycosylation, acylation, myristylation, palmitylation, farnesylation and processing) [S. cerevisiae, YKL201c] 2e-04
 30 [FUNCAT] e amino acid metabolism and transport [M. genitalium, MGD42] 4e-04
 [FUNCAT] 30.13 organization of chromosome structure [S. cerevisiae, YDR285w] 7e-04
 35 [FUNCAT] n secretion and adhesion [M. jannaschii, MJ0291] 0.001
 [FUNCAT] 05.04 translation (initiation, elongation and termination) [S. cerevisiae, YAL035w] 0.001
 [BLOCKS] BLO0326 Tropomyosins proteins
 40 [BLOCKS] PRO0545E
 [BLOCKS] PRO0041F
 [SCOP] d2tmab_1.105.4.1.1 Tropomyosin [rabbit (Oryctolagus cuniculus) 5e-05
 [EC] 3.6.1.32 Myosin ATPase 5e-16
 45 [PIRKW] nucleus 2e-35
 [PIRKW] phosphotransferase 5e-10
 [PIRKW] duplication 2e-09
 [PIRKW] citrulline 7e-09
 [PIRKW] tandem repeat 2e-13
 50 [PIRKW] heterodimer 2e-08
 [PIRKW] heart 2e-11
 [PIRKW] endocytosis 3e-10
 [PIRKW] polymorphism 1e-09
 [PIRKW] transmembrane protein 1e-12
 55 [PIRKW] serine/threonine-specific protein kinase 5e-10
 [PIRKW] cell wall 7e-09
 [PIRKW] zinc finger 3e-10
 [PIRKW] surface antigen 1e-08

	[PIRKW]	DNA binding 6e-12
	[PIRKW]	metal binding 3e-10
	[PIRKW]	muscle contraction 2e-13
	[PIRKW]	brain 8e-08
5	[PIRKW]	acetylated amino end 4e-09
	[PIRKW]	actin binding 5e-16
	[PIRKW]	endoplasmic reticulum 4e-09
	[PIRKW]	mitosis 3e-15
	[PIRKW]	microtubule binding 3e-15
10	[PIRKW]	ATP 5e-16
	[PIRKW]	chromosomal protein 2e-08
	[PIRKW]	receptor 4e-10
	[PIRKW]	thick filament 2e-13
	[PIRKW]	phosphoprotein 5e-16
15	[PIRKW]	glycoprotein 4e-10
	[PIRKW]	skeletal muscle 7e-11
	[PIRKW]	calcium binding 7e-09
	[PIRKW]	alternative splicing 3e-13
	[PIRKW]	DNA condensation 2e-08
20	[PIRKW]	coiled coil 5e-16
	[PIRKW]	P-loop 5e-16
	[PIRKW]	heptad repeat 3e-13
	[PIRKW]	methylated amino acid 2e-13
	[PIRKW]	basement membrane 1e-09
25	[PIRKW]	immunoglobulin receptor 2e-09
	[PIRKW]	peripheral membrane protein 3e-10
	[PIRKW]	cardiac muscle 2e-11
	[PIRKW]	extracellular matrix 1e-09
	[PIRKW]	hydrolase 5e-16
30	[PIRKW]	microtubule 1e-11
	[PIRKW]	muscle 1e-09
	[PIRKW]	membrane protein 1e-09
	[PIRKW]	EF hand 7e-09
	[PIRKW]	protein biosynthesis 4e-09
35	[PIRKW]	cytoskeleton 3e-13
	[PIRKW]	hair 7e-09
	[PIRKW]	Golgi apparatus 1e-11
	[PIRKW]	calmodulin binding 3e-10
	[SUPFAM]	myosin heavy chain 5e-16
40	[SUPFAM]	conserved hypothetical P115 protein 4e-10
	[SUPFAM]	IgA Fc receptor 7e-09
	[SUPFAM]	centromere protein E 3e-15
	[SUPFAM]	unassigned Ser/Thr or Tyr-specific protein kinases 5e-10
45	[SUPFAM]	calmodulin repeat homology 7e-09
	[SUPFAM]	myosin motor domain homology 5e-16
	[SUPFAM]	alpha-actinin actin-binding domain homology 5e-10
	[SUPFAM]	hypothetical protein MJ0914 4e-08
	[SUPFAM]	tropomyosin 6e-09
50	[SUPFAM]	plectin 5e-10
	[SUPFAM]	trichohyalin 7e-09
	[SUPFAM]	pleckstrin repeat homology 1e-08
	[SUPFAM]	ribosomal protein S10 homology 5e-10
	[SUPFAM]	giantin 4e-13
55	[SUPFAM]	protein kinase homology 5e-10
	[SUPFAM]	protein kinase C zinc-binding repeat homology 1e-08
	[SUPFAM]	kinesin motor domain homology 3e-15
	[SUPFAM]	human early endosome antigen 1 3e-10

-168-

5

10

15

20

25

30

(No Prosite data available for DKFZphamy2_2b19.2)
(No Pfam data available for DKFZphamy2_2b19.2)

DKFZphamy2_2c22

5 group: metabolism

DKFZphamy2_2c22 encodes a novel 364 amino acid protein with similarity to the 1-acyl-glycerol-3-phosphate acyltransferase of Zea mais.

10 It contains one leucine zipper. The protein is believed to play a role in fatty acid metabolism. It is ubiquitous expressed, with a slight predominance in uterus, placenta and foreskin.

15 The new protein can find application in modulation of fatty acid metabolism and as a new enzyme for biotechnological production processes.

20 weak similarity to 1-acyl-glycerol-3-phosphate acyltransferase (Zea mais)

~~perhaps complete cds.~~

25 Sequenced by MediGenomix

Locus: /map="8"

30 Insert length: 3403 bp
Poly A stretch at pos. 3373, polyadenylation signal at pos. 3351

```

    1 AGATGCTGCT GTCCCTGGTG CTCCACACGT ACTCCATGCG CTACCTGCTG
35  51 CCCAGCGTCG TGCTCTCTGGG CACGGCGCCC ACCTACGTGT TGGCCTGGGG
    101 GGTCTGGCGG CTGCTCTCCG CCTTCCTGCC CGCCCGCTTC TACCAAGCGC
    151 TGGACGACCG GCTCTACTGC GTCTACCAGA GCATGGTGCT CTTCTTCTTC
    201 GAGAATTACA CCGGGGTCCA GATATTGCTA TATGGAGATT TGCCAAAAAA
    251 TAAAGAAAAT ATAATATATT TAGCAAATCA TCAAAGCACA GTTGACTGGA
40  301 TTGTTGCTGA CATCTTGGCC ATCAGGCAGA ATGCGCTAGG ACATGTGCGC
    351 TACGTGCTGA AAGAAGGGTT AAAATGGCTG CCATTGTATG GGTGTTACTT
    401 TGCTCAGCAT GGAGGAATCT ATGTAAAGCG CAGTGCCAAA TTAAACGAGA
    451 AAGAGATGCG AAACAAGTTG CAGAGCTACG TGGACGCAGG AACTCCAATG
    501 TATCTTGCTG TTTTTCAGG AGGTACAAGG TATAATCCAG AGCAAACAAA
45  551 AGTCCTTTCA GCTAGTCAGG CATTGCTGC CCAACGTGGC CTTGCAGTAT
    601 TAAAACATGT GCTAACACCA CGAATAAAGG CAACTCACGT TGCTTTTGAT
    651 TGCATGAAGA ATTATTTAGA TGCAATTTAT GATGTTACGG TGGTTTATGA
    701 AGGGAAAGAC GATGGAGGGC AGCGAAGAGA GTCACCGACC ATGACGGAAT
    751 TTCTCTGCAA AGAATGTCCA AAAATTCTTA TTCACATTGA TCGTATCGAC
50  801 AAAAAAGATG TCCCAGAAGA ACAAGAACAT ATGAGAAGAT GGCTGCATGA
    851 ACGTTTCGAA ATCAAAGATA AGATGCTTAT AGAATTTTAT GAGTCACCAG
    901 ATCCAGAAAG AAGAAAAAGA TTTCTTGGGA AAAGTGTTAA TTCCAAATTA
    951 AGTATCAAGA AGACTTTACC ATCAATGTTG ATCTTAAGTG GTTTGACTGC
1001 AGGCATGCTT ATGACCGATG CTGGAAGGAA GCTGTATGTG AACACCTGGA
55 1051 TATATGGAAC CCTACTTGGC TGCCTGTGGG TTAATAATTA AGCATAGACA
    1101 AGTAGCTGTC TCCAGACAGT GGGATGTGCT ACATTGTCTA TTTTGGCGG
    1151 CTGCACATGA CATCAAATTG TTTCTGAAT TTATTAAGGA GTGTAAATAA
    1201 AGCCTTGTTG ATTGAAGATT GGATAATAGA ATTTGTGACG AAAGCTGATA
```

1251 TGCAATGGTC TTGGGCAAAC ATACCTGGTT GTACAACTTT AGCATCGGGG
1301 CTGCTGGAAG GGTAAAAGCT AAATGGAGTT TCTCCTGCTC TGTCCTTTTC
1351 CTATGAACTA ATGACAACCT GAGAAGGCTG GGAGGATTGT GTATTTTGCA
1401 AGTCAGATGG CTGCATTTTT GAGCATTAAT TTGCAGCGTA TTTCACTTTT
5 1451 TCTGTTATTT TCAATTTATT ACAACTTGAC AGCTCCAAGC TCTTATTACT
1501 AAAGTATTTA GTATCTTGCA GCTAGTTAAT ATTTTCATCTT TTGCTTATTT
1551 CTACAAGTCA GTGAAATAAA TTGTATTTAG GAAGTGTGAG GATGTTCAAA
1601 GGAAAGGGTA AAAAGTGTTT ATGGGGAAAA AGCTCTGTTT AGCACATGAT
1651 TTTATTGTAT TGCCTTATTA GCTGATTTTA CTCATTTTAT ATTTGCAAAA
10 1701 TAAATTTCTA ATATTTATTG AAATTGCTTA ATTTGCACAC CCTGTACACA
1751 CAGAAAAATGG TATAAAATAT GAGAACGAAG TTTTAAATTG TGACTCTGAT
1801 TCATTATAGC AGAACTTTAA ATTTCCCAGC TTTTGAAGA TTTAAGCTAC
1851 GCTATTAGTA CTTCCCTTTG TCTGTGCCAT AAGTGCTTGA AAACGTTAAG
1901 GTTTTCTGTT TTGTTTGTG TTTTAAATAT CAAAAGAGTC GGTGTGAACC
15 1951 TTGGTTGGAC CCCAAGTTCA CAAGATTTTT AAGGTGATGA GAGCCTGCAG
2001 ACATTCTGCC TAGATTTACT AGCGTGTGCC TTTTGCCTGC TTCTCTTTGA
2051 TTTCACAGAA TATTCATTCA GAAGTCGCGT TTCTGTAGTG TGGTGGATTG
2101 CCACTGGGCT CTGGTCCTTC CCTTGGATCC CGTCAGTGGT GCTGCTCAGC
2151 GGCTTGACAG CAGACTTGCT AGGAAGAAAT GCAGAGCCAG CCTGTGCTGC
20 2201 CCACCTTCAG AGTTGAACTC TTTAAGCCCT TGTGAGTGGG CTTCACAGC
2251 TACTGCAGAG GCATTTTGCA TTTGTCTGTG TCAAGAAGTT CACCTTCTCA
2301 AGCCAGTGAA ATACAGACTT AATTTGTCAT GACTGAACGA ATTTGTTTAT
2351 TTCCCATTAG GTTTAGTGGA GCTACACATT AATATGTATC GCCTTAGAGC
2401 AAGAGCTGTG TTCCAGGAAC CAGATCACGA TTTTTAGCCA TGGAACAATA
25 2451 TATCCCATGG GAGAAGACCT TTCAGTGTGA ACTGTTCTAT TTTTGTGTTA
2501 TAATTTAAAC TTCGATTTCC TCATAGTCCT TTAAGTTGAC ATTTCTGCTT
2551 ACTGCTACTG GATTTTGTCT GCAGAAATAT ATCAGTGGCC CACATTAAAC
2601 ATACCAGTTG GATCATGATA AGCAAAATGA AAGAAATAAT GATTAAGGGA
2651 AAATTAAGTG ACTGTGTTAC ACTGCTTCTC CCATGCCAGA GAATAAACTC
30 2701 TTTCAAGCAT CATCTTTGAA GAGTCGTGTG GTGTGAATTG GTTTGTGTAC
2751 ATTAGAATGT ATGCACACAT CCATGGACAC TCAGGATATA GTTGGCCTAA
2801 TAATCGGGGC ATGGGTAAAA CTTATGAAAA TTTCTCATG CTGAATTGTA
2851 ATTTTCTCTT ACCTGTAAAG TAAAATTTAG ATCAATTCCA TGCTTTGTGTT
2901 AAGTACAGGG ATTTAATATA TTTTGAATAT AATGGGTATG TTCTAAATTT
35 2951 GAACTTTGAG AGGCAATACT GTTGGGAATTA TGTGGATTCT AACTCATTTT
3001 AACCAAGGTAG CCTGACCTGC ATAAGATCAC TTGAATGTTA GGTTCATAG
3051 AACTATACTA ATCTTCTCAC AAAAGGTCTA TAAAATACAG TCGTTGAAAA
3101 AAATTTTGTA TCAAAATGTT TGGAAAATTA GAAGCTTCTC CTTAACCTGT
3151 ATTGATACTG ACTTGAATTA TTTTCTAAAA TTAAGAGCCG TATACCTACC
40 3201 TGTAAGTCTT TTCACATATC ATTTAAACTT TTGTTTGTAT TATTACTGAT
3251 TTACAGCTTA GTTATTAATT TTTCTTTATA AGAATGCCGT CGATGTGCAT
3301 GCTTTTATGT TTTTCAGAAA AGGGTGTGTT TGGATGAAAG TAAAAAATAA
3351 AAATAAAATC TTTCACTGTC TCTAAAAAAA AAAGAAAAAA AAAAAAATAA
3401 AAA

BLAST Results

50 No BLAST result

Medline entries

55 No Medline entry

Peptide information for frame 3

5 ORF from 3 bp to 1094 bp; peptide length: 364
 Category: similarity to known protein
 Classification: Metabolism
 Prosite motifs: LEUCINE_ZIPPER (105-126)

10

1 MLLSLVLHTY SMRYLLPSVV LLGTAPTYVL AWGVWRLLSA FLPARFYQAL
 51 DDRLYCVYQS MVLFFFENYT GVQILLYGDL PKNKENIIYL ANHQSTVDWI
 101 VADILAIRQN ALGHVRYVLK EGLKWLPYLG CYFAQHGGIY VKRSKAFNEK
 151 EMRNKLQSYV DAGTPMYLVI FPEGTRYNPE QTKVLSASQA FAAQRGLAVL
 15 201 KHVLTPIKA THVAFDCMKN YLDAIYDVT VYEGKDDGGQ RRESPTMTEF
 251 LCKECPKIHI HIDRIDKKDV PEEQEHMRRW LHERFEIKDK MLIEFYESP
 301 PERRKRFPKG SVNSKLSIKK TLPSMLILSG LTAGMLMTDA GRKLYVNTWI
 351 YGTLLGCLWV TIKA

20

BLASTP hits

No BLASTP hits available

25

Alert BLASTP hits for DKFZphamy2_2c22, frame 3

No Alert BLASTP hits found

30

Pedant information for DKFZphamy2_2c22, frame 3

Report for DKFZphamy2_2c22.3

35

[[LENGTH]] 364
 [[MW]] 42072.47
 [[pI]] 9.18
 [[HOMOL]] TREMBL:CEAF3136_1 gene: "F28B3.5"; Caenorhabditis
 40 elegans cosmid F28B3. 2e-36
 [[FUNCAT]] 99 unclassified proteins [[S. cerevisiae, YDR018c]]
 7e-13
 [[FUNCAT]] 01.06.01 lipid, fatty-acid and sterol biosynthesis
 [[S. cerevisiae, YDL052c]] 4e-05
 45 [[FUNCAT]] 30.99 other cellular organization [[S. cerevisiae,
 YDL052c]] 4e-05
 [[BLOCKS]] BLO1263A
 [[BLOCKS]] BP00989A
 [[PIRKW]] transmembrane protein 2e-11
 50 [[SUPFAM]] probable membrane protein YBR042c 2e-11
 [[PROSITE]] LEUCINE_ZIPPER 1
 [[KW]] Alpha_Beta
 [[KW]] LOW_COMPLEXITY 3.57 %

55

SEQ MLLSLVLHTYSMRYLLPSVVLLGTAPTYVLAWGVWRLLSAFLPARFYQALDDRLYCVYQS
 SEG
 PRD ccchhhhhhhhhccccccccccccccccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh

5 SEQ MVLFFFENYTGVDILLYGDLPKNKENIIYLANHQSTVDWIVADILAIRQNALGHVRYVLK
SEG
PRD hhhhhhhceeeeeeeeeecccccccccceeeeccccchhhhhhhhhhhhhcccccchhhhhh

10 SEQ EGLKWLPYGCYFAQHGGIYVKRSKAFNEKEMRNKLQSYVDAGTPMYLVIFPEGTRYNPE
SEG
PRD hhhccccccccceeeccceeeecccccchhhhhhhhhhhcccccceeeecccccchhh

15 SEQ RRESPTMTEFLCKECPKIHIDRIDKKDVPPEQEHMRRWLHERFEIKDKMLIEFYESP
SEGxxxxxxxxxxxxx.....
PRD cccccchhhhhccccceeeecccccccccccccchhhhhhhhhhhhhhhhhhhhhhhccc

20 SEQ PERRKRFPGKSVNSKLSIKKTLPSMLILSGLTAGMLMTDAGRKLYVNTWIIYGTLLGCLWV
SEG
PRD cccccccccccchhhhhhhchhhhhchhhhhhhhhhhccccceeeeeecehhhhhhhh

25 SEQ TIKA
SEG
PRD hccc

Prosites for DKFZphamy2_2c22.3

30 PSD00029 105->127 LEUCINE_ZIPPER PD0C00029

(No Pfam data available for DKFZphamy2_2c22.3)

DKFZphamy2_2f18

5 group: signal transduction

DKFZphamy2_2f18 encodes a novel 215 amino acid protein with similarity to sodium channel protein beta1 of *Rattus norvegicus*.

10 The sodium channel protein beta 1 of *Rattus norvegicus* is crucial in the assembly, expression, and functional modulation of the heterotrimeric complex of the rat brain sodium channel. The expression of the new protein seems to be restricted to brain, all matching ESTs isolated so far, derive from there.

15 The new protein can find application in modulating the sodium channel beta 1, studying the expression profile in neurodegenerative diseases and of amygdala-specific genes.

20

similarity to sodium channel protein beta1 (*Rattus norvegicus*)

Pedant: SIGNAL_PEPTIDE

25

Sequenced by MediGenomix

Locus: unknown

30 Insert length: 4052 bp

Poly A stretch at pos. 4035, no polyadenylation signal found

```

35      1 CAGGGCTGAC AGCACACACG GCCTGGGGGC CTAGAGAAGG ATTGCTGATC
      51 ACCTGCCACC CAGGGTCGGG GCCCGCACC ATCCGGGGGC GAGCTCCCGG
     101 GAAGGGGCTC CCCCTCTACA CCCACCCCC AACCTCTGAC ATCGCCGGCC
     151 GAACGGGAGC TGCCGCTTCC TCCCCGGCCC CGCTGCACCT CCCCAGGGAG
     201 CCGAGGGCGG GCGTGGACGG GACCGACGTG GAACGCATTG TGTAGCCAG
     251 ACGGGCGGCC CCGGCGGCTT CGGGAGTGGG GTCACGCCCA GCTGGAGAAG
     40 301 CAGTTAGGGC GGACGAAGCA GGAGCCGCGG GGCTGGGAGG ATTCCAGTCG
     351 GAACGCAACC GATCCTGGGG AGGCGAGAGG TGAATCAACC TGGACCCCTC
     401 CACAGCCTGG CTGCTAGGCC AGCAGTGCGA CTCCCTTCCG AGCTGAGCTT
     451 ACCCTGGGCG CAAACGAGCG AGGCAGGGGC GCGAGTGGAA GCTGGAGTTC
     501 CGGGGTGGGC GGGGAGGCGA CTGTCCGTGG TGCTGAGCGC CGGCGAGAGC
     45 551 GGGCGCGGAG CGGCTGATCA GCTCCCTCGA ACTGGGGAGG TCCAGTGGGG
     601 TCGCTTAGGG CCCAAAGCCC CCGCCCGGCT CCAAAAGCTC CCAGGGCCTC
     651 CCCAGGCACC GGTGCTCGGC CCTTCCTTCG GTCAGAAAGT CGCCCCCTGG
     701 GGGCAGTTCG TCCCAAAGGG TTTCCTCGAA AGAATCTGAG AGGGCGCAGT
     751 CCTTGACCGA GGAATCTCT CTGTGTAGCC TTGGAAGCCG CCAGCCCCAG
     50 801 AAGATGCCTG CCTTCAATAG ATTGTTTCCC CTGGCTTCTC TCGTGCTTAT
     851 CTA CTGGGTC AGTGTCTGCT TCCCTGTGTG TGTGGAAGTG CCCTCGGAGA
     901 CGGAGGCCGT GCAGGGCAAC CCCATGAAGC TCGCTGCAT CTCTGCATG
     951 AAGAGAGAGG AGGTGGAGGC CACCACGGTG GTGGAATGGT TCTACAGGCC
    1001 CGAGGGCGGT AAAGATTTCC TTATTTACGA GTATCGGAAT GGCCACCAGG
    55 1051 AGGTGGAGAG CCCCTTTTCA GGGCGCCTGC AGTGGAAATG GAGCAAGGAC
    1101 CTGCAGGACG TGTCCATCAC TGTGCTCAAC GTCACTCTGA ACGACTCTGG
    1151 CCTCTACACC TGCAATGTGT CCCGGGAGTT TGAGTTTGAG GCGCATCGGC
    1201 CCTTTGTGAA GACGACGCGG CTGATCCCCC TAAGAGTCAC CGAGGAGGCT

```

1251 GGAGAGGACT TCACCTCTGT GGTCTCAGAA ATCATGATGT ACATCCTTCT
1301 GGTCTTCTCT ACCTTGTGGC TGCTCATCGA GATGATATAT TGCTACAGAA
1351 AGGTCTCAAA AGCCGAAGAG GCAGCCCAAG AAAACGCGTC TGACTACCTT
1401 GCCATCCCAT CTGAGAACAA GGAGAACTCT GCGGTACCAG TGGAGGAATA
5 1451 GAACAGGAGC AGTGTGACAT GAGGTGGCCT GAACACCTGA GGGACTGGAC
1501 ATCCCATGTT CAGCAATGTC AATGGCATCA GGAGGGCGCC CCAAGGGCCC
1551 CATCGCTTCC CTTCATGCAT CCATTGTTCT GTTCATTCTAT TCATCCATAC
1601 ATCCACCTGC CTCTGAGCTT TCACCTCTGA CTCCCTAACT CCATCAGACC
1651 TCTACGCACC ATAAGACTCT GCCAGAAGTGA AGAAGCCAAC ATTTCTACAT
10 1701 AGACTCAACC TCACCCTCTC CTAGTTTTTC AACAAGACAC TCCAAAGCCA
1751 ACTGGATTTT TCCCCTGTGC TCCAAATGAC TTTGTACAAG TGCTGGAGTT
1801 AGCACCTCCC TCTGCCCTTA ACTGGCTGGA ACTGGTTCAT TCTCCATTAC
1851 TGCAAGAGAA TGGAAGTCTT AATAGAAGGA AGCAGGAGTG ATTAGTTCGG
1901 GTTAAAGCAA AAGTGTGTCA TGAAGTTGGA TTCCCTGAAG TCAGTTTTGT
15 1951 CAGGTTTCATG GCCCACCTTG CTACAGCATC AGAGTGAAGC ACGCTGTCT
2001 AGGTTCTCCA GTGACAGAAA GATCCTGAAG CATGGACTAA CATGCTCTCT
2051 GGAGCTTAGT ACTCCAGAGC TAGATCCTGA TGGGTCTCTA AGGTTCCCTC
2101 CAAGAAGACA AGGACAGGAG ACTTGGGAAG GACCAATGGT AATTTAAGTG
2151 GCTCTTAAAA AGTCATGCAA TATGTTTTCTG GACACGTTCC TGATCCTATT
20 2201 GCGATAATGT ATGTGTGCCC TCCCTGTGGG CACACCACCT GGGCATTAGG
2251 ACTGAAATTC CTGAGTTCTT CCTCTCAAAA TTTCTGTGCA CCGATTATAT
2301 TCCTCATTTT ACATACAGGA GGCAACTAAG ACTCATACAG GGTCAACTG
2351 AATAAGAGGC TTAAGAGGAT AAAGTGGAGC AGAAATAAGC CTTAGGTGCT
2401 GCCCAGTTTA CACTTCCTGG GATGGATGTT TTTGTTTGTG TTGTTTTTTG
25 2451 TTTTTTTTGT TTGAGATGGA GTCTCACTCT GTCACCTAGG CTAGAGTGCA
2501 GTGGTGTGAT CTCGGCTCAC TGCAACCTCT GCCTCTTGGG TTCAAGCAAT
2551 TCTCATGCCT CGGCCTCTCC AGTAGCTGGG ATTACAGGTG TGCAACCACCA
2601 CGCTGGCTA AATTTTGTAT TTTTAGTACA GACAGGGTTT GACTATGTTG
2651 GCCAGGCTAG TCTTGAAGTC CTGACCTCAA ATGACCCACC GAGCTCAGCC
30 2701 TCCCAAGTG CTGAGATTAC AGGCGTGAGG CACTGCGCCC GGTGGATAAC
2751 TTTGTTTCTG AAAAGACTGA CATTGAACTT GTCTATGGCA ATGCTTCTTT
2801 CACAAGCACG GACTGGGCTG AGGTCAACTC TGATAGATTG AGATGACTAG
2851 AAATTGGCCA AAAAAGCAGG GAGAAGAACA TGAGGTAGAC TTAAAGAACT
2901 TCCTTTATGT AAAGATCTGT GACTCTGAAA TATCCTCCAA AAGGAGAGTG
35 2951 CATCTGAGAC TGATATTTAA ACTAAGAAAA ATGTTTAGTC TGAGATGGAT
3001 CATAAGTAAA TGAGCAGTGT GAGAGGGGAG GGATGGGTAG GTGCTTTCCA
3051 AATACTTCGC CTATGAATGC ATAATTTTCA GATTTTTTTC CCTAGATTT
3101 TGAGGGAGCA GAGAACTGG AAAAACTTT AGTCAATATC TCGTGTTCAT
3151 TTTTAATTAA GTGACAGGTC CAAGTGTGAC ATCCTTCAGC ACCCAGGGAC
40 3201 AAGAGAGGGG AAAGATGCTT TATGGAATGT AAGAAGATGA AGGTGACTGG
3251 GATTACGCGA GAGAGAGGTC CCTCAGACCT GGGACCTCCC TTTATAGGGA
3301 AAGACCATAT TCCATAGGTT TAGGGCTTTA CCTTAAAAGC TCATTTTTTT
3351 CATTCTTCCA TCCCTAGGAA AGTACTTAAA ACCAGACTTT TAAATTTTTA
3401 TTTATTTATT ATTTATTTTT TGAGACAGAT TCTCACTCTG TCTCCAGGC
45 3451 TAGAGTGCAG TGGTGCAATC TCAGCTCACT GCAGCCTCAA CTGCCCCAGG
3501 TTAAAGCAAT CCTCCCACCT CAGCCCCAG GTAAGTGGGA CTACAGGCAT
3551 GCACCACCAT GCCTGGCTAA TTTTTGTATT TTATGTAGAG ACAGGGGTCT
3601 TGCCATGTCT CCCAGGCTGA TCTTGAAGTC CTGGGCTCAA GCAATCTGCC
3651 AGCCTCAGCC TCTCAAAGTG CTGGGATTAC AGGCCTGAGC AACTGTGCCT
50 3701 GGCCCAAAAC CAGACCGTTA ACACATTTAA GAGTCTGATT TTGTTGAAGA
3751 AAATATTTGC AATAAATTCA AGACTCTTCT TATTGGTAAT TTTCCACACA
3801 ATCCCTCTGA AATAAGGGAG AGGATATAGA CCTTTTAAAC TTTATAGTTA
3851 GAAAAATTGG CCTCAGTGTG AAATTTTTCC AGTCCCATAG CTCATTGGATG
3901 GCACCAGCTT GCGGTAGTAG CAAGATGCTT ACTACCACAC CGTTTTCTC
55 3951 GGTGGCCCAA TAGCTCGTGT ATCTAAGTTG AACCCGGCAG TATGCTATGAT
4001 TGCTTTTTTC TCTTCTTTTT AAAAAACCC AACTCAAAAA AAAAAAATAA
4051 AA

BLAST Results

5 No BLAST result

Medline entries

- 10 92271207:
Isom LL, De Jongh KS, Patton DE, Reber BF, Offord J, Charbonneau
H,
Walsh K, Goldin AL, Catterall
15 WA.; Primary structure and functional expression of the beta 1
subunit
of
the rat brain sodium channel. Science 1992 May 8;256(5058):839-42
- 20 96235151:
Belcher SM, Howe JR.; Cloning of the cDNA encoding the sodium
channel
beta 1 subunit from rabbit. Gene 1996 May 8;170(2):285-6
- 25 93357746:
McClatchey AI, Cannon SC, Slaugenhaupt SA, Gusella JF.; The
cloning and
expression of a sodium channel beta
1-subunit cDNA from human brain. Hum Mol Genet 1993 Jun;2(6):745-
30 9

Peptide information for frame 3

35

ORF from 804 bp to 1448 bp; peptide length: 215
Category: similarity to known protein
40 Classification: Transmembrane proteins unclassified

1 MPAFNRLFPL ASLVLIYWVS VCFPVCVEVP SETEAVQGNP MKLRCISCMK
51 REEVEATTVV EWFYRPEGGK DFLIYEYRNG HQEVESPFQG RLQWNGSKDL
101 QDVSITVLNV TLNDSGLYTC NVSREFEFEA HRPFVKTTTL IPLRVTEEAG
45 151 EDFTSVVSEI MMYILLVFLT LULLIEMIYC YRKVSKAEEA AQENASDYLA
201 IPSENKENSA VPVEE

BLASTP hits

50

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_2f18, frame 3

55

PIR:JC4788 sodium channel protein beta1 chain - rabbit, N = 1,
Score =
434, P = 8.3e-41

PIR:A55734 sodium channel, voltage-gated, beta-1 chain precursor
human, N = 1, Score = 428, P = 3.6e-40

PIR:A42737 sodium channel beta 1 subunit - rat, N = 1, Score =
429, P =
2.8e-40

>PIR:JC4788 sodium channel protein beta1 chain - rabbit
Length = 218

HSPs:

Score = 434 (65.1 bits), Expect = 8.3e-41, P = 8.3e-41
Identities = 100/214 (46%), Positives = 129/214 (60%)

Query: 10

LASLVLIYWPVSVCFVCEVPSETEAVQGNPMKLRCSMKREEVEATTVVEWVYRPEGG 69
LA +V VS + CEEV SETEAV G K+ CISC +R E A T

EW +R +G

Sbjct: 5

LAFVVGAAALVSSAWGGCCEVDSETEAVYGMTFKILCISCKRRSETTAETFTETWTFRQKGT 64

Query: 70 KDFL-IYEYRNGHDEVESP--FQGRLEWNGS---

KDLQDVSIITVLNVTLNDSGLYTCNVS 123

++F+ I Y N ++E F+GR+ WNGS KDLQD+SI + NVT N

SG Y C+V

Sbjct: 65

EEFVKILRYENEVLQLEEDERFEGRVVWNGSRGTDLQDLSIFITNVTYNHSGDYQCHVY 124

Query: 124

REFEFEAHRPFVKTTTRLIPLRVTEEAGEDFTSVVSEIMMYIXXXXXXXXXXXIEMIYCYRK 183

R FE + - + I L V ++A D S+VSEIMMY+

EM+YCY+K

Sbjct: 125

RLLSFENYEHNTSVVKKIHLEVVDKANRDMASIVSEIMMYVLIVVLTITWLVAEMVYCYKK 184

Query: 184 VSKAEAAA-QENASDYLAIPSENKEN-SAVPVEE 215

++ A EAA QENAS+YLAI SE+KEN + V V E

Sbjct: 185 IAAATEAAAQENASEYLAITSESKENCTGVQVAE 218

Pedant information for DKFZphamy2_2f18, frame 3

Report for DKFZphamy2_2f18.3

[[LENGTH]] 215

[[MW]] 24702.40

[[pI]] 4.69

[[HOMOL]] PIR:JC4788 sodium channel protein beta1 chain -
rabbit 3e-41

[[BLOCKS]] BL004010 Prokaryotic sulfate-binding proteins

[[BLOCKS]] BP00570

```

[SCOP]      dlnu__ 2.1.1.1.1 Myelin membrane adhesion
molecule PO [ra 2e-43]
[PIRKW]      Schwann cell 2e-07
[PIRKW]      transmembrane protein 1e-40
5 [PIRKW]      myelin 2e-07
[PIRKW]      phosphoprotein 5e-07
[PIRKW]      glycoprotein 1e-40
[PIRKW]      structural protein 2e-07
[PIRKW]      muscle 1e-40
10 [PIRKW]      membrane protein 5e-07
[SUPFAM]     immunoglobulin homology 2e-07
[SUPFAM]     myelin PO protein 2e-07
[PFAM]       IG (immunoglobulin) superfamily
[KW]         All_Beta
15 [KW]         3D
[KW]         SIGNAL_PEPTIDE 23
[KW]         LOW_COMPLEXITY      4.65 %

20 SEQ  MPAFNRLFPLASLVLIYWVSVCFPVCVEVPSETEAVQGNPMKLCISCMKREEVEATTVV
SEG  .....
lnu- .....CEEEECCEEEETTTbCEEECE-
EEEECCCCCCCCCEE

25 SEQ  EWFYRPEGGKDFLIYEYRNGHQEVESPFQGRLLQWNGSKDLQDVSITVLNVTNLNDSGLYTC
SEG  .....
lnu- .....
EEEEETTTCCCEEEEEETTEEEETTTTTTTEEECCBGGGCBCEEECCbTTTTTEEEEE

30 SEQ  NVSREFEFEAHRPFVKTTRLIPLRVTEEAGEDFTSVVSEIMMYILLVFLTLWLLIEMIYC
SEG  .....xxxxxxxxxxxx.....
lnu- .....
EE.....

35 SEQ  YRKVSKAEAAQENASDYLAIPSENKENSAPVVEE
SEG  .....
lnu- .....

40 (No Prosite data available for DKFZphamy2_2f18.3)

                                Pfam for DKFZphamy2_2f18.3

45 HMM_NAME  IG (immunoglobulin) superfamily

HMM
*yrNgqpipssegyWytRweqqgRYsisifqLtIisWepeDsGtYWcmV*
50 YRNG ++ E+ ++ R++++G ++ +++T+ +++ +DSG
Y+C+V
Query . 77 YRNGHQEV--
ESPFQGRLLQWNGSKDLQDVSITVLNVTNLNDSGLYTCNV 122

55

```

DKFZphamy2_2f22

5 group: nucleic acid management

DKFZphamy2_2f22 encodes a novel 479 amino acid protein with similarity to YDL153c of *Saccharomyces cerevisia*.

10 The novel protein is ubiquitously expressed. YDL153c is involved in transcriptional silencing.

The new protein can find application in modulation of transcription, e.g. transcriptional silencing.

15

putative protein

probably complete cds.

20 perhaps differential polyadenylation
YDL153c is involved in transcriptional silencing

Sequenced by MediGenomix

25 Locus: /map="4"

Insert length: 2019 bp

Poly A stretch at pos. 2000, polyadenylation signal at pos. 1981

30

```
1 GGAGTCTGCA AACTCCGGTG GTAGGGGAGC GCGCTGCTGT TTAGAGCCAC
51 GAGTTACCGG AGCGCCTGAT TCCTGCGCCG AAGTCAAGTG TGGCCGAAAG
101 TCCGGAGTCG CTGTAAACC TGAGATTGTG AGCCATGGTG GGGAGATCCC
151 GCGGGCGCGG AGCAGCTAAG TGGGCAGCTG TCGAGCCAA GGCAGGTCCC
35 201 ACGCTACCG ACGAAAATGG AGATGATTTA GGATTGCCAC CCTCACCAGG
251 GGACACCAGC TACTACCAAG ATCAGGTAGA TGACTTTCAT GAGGCACGAT
301 CCCGGGCCGC CTTAGCTAAG GGCTGGAATG AAGTACAGAG TGGAGACGAG
351 GAGGATGGCG AGGAGGAGGA GGAGGAGGTG CTAGCCCTAG ATATGGACGA
40 401 TGAGGACGAC GAAGATGGAG GGAATGCGGG GGAGGAGGAG GAGGAGGAGA
451 ATGCCGATGA TGATGGTGGG AGCTCCGTGC AAAGTGAAGC TGAGGCCTCT
501 GTGGATCCCA GTTTGTCTGT GGGTCAGAGG AAAAACTTT ACTATGACAC
551 GGACTATGGT TCCAAGTCCC GAGGCCGGCA GAGTCAACAG GAGGCAGAGG
601 AGGAGGAAAG AGAGGAGGAG GAGGAGGCAC AGATCATTCA GCGGCGCCTA
651 GCCCAAGCGC TGCAAGAGGA TGATTTTGGT GTCGCCTGGG TTGAGGCCTT
45 701 TGCAAAACCA GTGCCTCAGG TAGATGAGGC TGAGACACGG GTCGTGAAGG
751 ATTTGGCTAA AGTTTCAGTG AAAGAGAAGC TGAAAATGTT GCGAAAGGAA
801 TCACCAGAAC TCTTGGAGCT GATAGAAGAC CTGAAAGTCA AGTTGACAGA
851 GGTTAAGGAT GAGCTGGAGC CATTGTTAGA GTTGGTGGAA CAAGGGATCA
901 TTCCACCCGG AAAAGGAAGC CAATACTTGA GGACCAAGTA CAACCTCTAC
50 951 TTGAATTATT GCTCGAACAT CAGTTTTTAT TTGATCCTGA AAGCTAGGAG
1001 AGTCCCAGCA CATGGACATC CTGTCATAGA AAGGCTTGTT ACCTACCGAA
1051 ATTTGATCAA CAAGCTGTCC GTTGTGGATC AGAAGCTGTC CTCAGAAATT
1101 CGTCATCTGT TGACACTTAA GGATGATGCT GTAAAGAAAG AACTGATTCC
1151 AAAAGCAAAA TCCACCAAGC CCAAACCAA GTCTGTTTCA AAGACTTCTG
55 1201 CTGCTGCCTG TGCTGTTACA GATCTTCTG ATGATTCTGA TTTTGATGAA
1251 AAAGCAAAAC TGAAGTACTA TAAAGAAATA GAAGACAGGC AAAAGCTAAA
1301 GAGAAAGAAA GAAGAAAATA GCACTGAAGA ACAGGCTCTT GAAGATCAAA
1351 ATGCAAAAGAG AGCTATTACC TATCAAATTG CTAAAAATAG GGGACTTACT
```

```
1401 CCTAGGAGAA AGAAGATTGA TCGCAATCCC AGAGTGAAAC ACAGAGAGAA
1451 GTTCAGAAGA GCCAAAATTA GAAGAAGAGG CCAGGTTTCGT GAAGTTCGTA
1501 AAGAAGAGCA ACGTTATAGT GGTGAATTAT CTGGCATTCTG TGCAGGAGTT
1551 AAAAAGAGCA TTAAGCTTAA ATGAAGTTTT TGCTTAGCAT AAGGTTTTTG
5 1601 GCAGTTTTTG ATCAATAAAT TTTTACTTTT AACTAAAGTC ATTGTATTAA
1651 TATATAATAC TTTAAATTTT AAAAATTCTT GTCCACAAGG AAATTTGTCT
1701 GGGTTATTGG ACAATTTATA AGAACTATGG GAGCAATATG AAGGTGCTTG
1751 AGAAAAGAGA TGATGTTGAA GTTTTCCAAT ATTCTGTTGA AGTTTTCCAA
1801 TATTAAGTAT TAGCTTAGGG AAATTTTACA GTTCATTGTG GAGTGTAAAA
10 1851 CTTAGAACAT GTGTAACCTT TCACATAAAG AGAATGCATC TTTGACAGTT
1901 ATCTTATTTG TAAGGCAGCC TATAAAATAG TTCTGAAGTA TTTTATTTAC
1951 CTAACATAAA TTATTGGGCC AGATACTTGT TAATAAATGG GCTTAATGTC
2001 AAAAAAAAAA AAAAAAAAAA
```

15

BLAST Results

No BLAST result

20

Medline entries

25 No Medline entry

Peptide information for frame 3

30

ORF from 135 bp to 1571 bp; peptide length: 479
Category: similarity to unknown protein
Classification: Nucleic acid management

35

```
1 MVGRSRRRGÁ AKWAAVRKA GPTLTDENGD DLGLPPSPGD TSYQQDQVDD
51 FHEARSRAAL AKGWNEVQSG DEEDGEEEE EVLALDMDE DDEGGNAGE
101 EEEENADDD GGSSVQSEAE ASVDPSLSWG QRKLYYDTD YGSKSRGRQS
151 QQEAEERE EEEEAQIIQR RLAQALQEDD FGVAVVEAFA KPVPQVDEAE
40 201 TRVVKDLAKV SVKEKLMLR KESPELLELI EDLKVKLTEV KDELEPLLEL
251 VEQGIIPP GK GSQYLRTKYN LYLNYCSNIS FYLILKARRV PAHGHPVIER
301 LVTYRNLINK LSVVDQKLSS EIRHLLTLKD DAVKKELIPK AKSTKPKPKS
351 VSKTSAAACA VTDLSDDSD F DEKAKLKYK EIEDRQKLKR KKEENSTEEQ
401 ALEDQNAKRA ITYQIAKNRG LTPRRKKIDR NPRVKHREKF RRAKIRRRGQ
45 451 VREVRKEEQR YSGELSGIRA GVKKSIKIK
```

BLASTP hits

50

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_2f22, frame 3

55 PIR:S67701 hypothetical protein YDL153c - yeast (Saccharomyces cerevisiae), N = 4, Score = 134, P = 1.8e-08

PIR:T08694 hypothetical protein DKFZp5640092.1 - human
(fragment), N =
1, Score = 141, P = 5.8e-07

5 TREMBL:SPBC3B8_9 gene: "SPBC3B8.09"; product: "hypothetical
protein";
S.pombe chromosome II cosmid c3B8., N = 2, Score = 164, P = 6.2e-
13

10 >TREMBL:SPBC3B8_9 gene: "SPBC3B8.09"; product: "hypothetical
protein";
S.pombe chromosome II cosmid c3B8.
Length = 597

15 HSPs:

Score = 164 (24.6 bits), Expect = 6.2e-13, Sum P(2) = 6.2e-13
Identities = 44/126 (34%), Positives = 68/126 (53%)

20 Query: 367 DSDFDKAKLKYYKEIEDRQKLKRK-KEEN-----STEEQALE-
DQNAKRAITYQ 414
D + +++ L YY+ ++ + K+ +K ++EN S + +E +
KR IT
25 Sbjct: 472
DREVEDQDDLDDYYESLDKKSMAKKLRKENHDLERDLIRASRHP ELIELGEGDKRGITLD 531

Query: 415 IAKNRGLTPRRKKIDRNPRVKHXXXXXXXXXXXXXGQVREVRKEEQR-
YSGELSGIRAGVK 473
30 IAKNRGLTPRR K +RNPR+K + + Q Y+GE
+GI+AG+
Sbjct: 532
IAKNRGLTPRRPKENRNPRLLKKRMRYEKAKKKLASKKAIYKGAPQGGYAGEQTGIKAGLV 591

35 Query: 474 KSIK LK 479
KSIKL+
Sbjct: 592 KSIKLQ 597

Score = 80 (12.0 bits), Expect = 6.2e-13, Sum P(2) = 6.2e-13
40 Identities = 29/129 (22%), Positives = 66/129 (51%)

Query: 197 DEAE TRVVK-DLAKVSVKEKLKMLRKESP--ELLELIE----
DLKVKL TEVKDELEPLLE 249
D ++ + +K D + +++E ++ + + P ELL+++E + ++ L E+
45 ++L+P L
Sbjct: 173 DNSDLKSIKQDSSAAAEELVQQISPDLPRTELLKILEAKHPEFQLFLDEL-
NQLKPQLN 231

Query: 250 LVEQGIIPP GKG S QYLR TKYNLYLNYCSNISFY L-
50 ILKARRVPAHGHPVIERLV TYRNLI 308
+++ + SQ L+ + Y S ++FY +LK HP++
LV +
Sbjct: 232 EIKEKL-
KTYPSSQLLQAACTALSTYISFLT FYFALLKDGEEDLKNHPIMVDLVRCKQ TW 290

55 Query: 309 NKLSVVDQKLS 319
+D+ L+
Sbjct: 291 ESYCGLDEVLT 301

Score = 59 (8.9 bits), Expect = 9.2e-11, Sum P(2) = 9.2e-11
Identities = 18/59 (30%), Positives = 35/59 (59%)

5 Query: 196 VDEAETRVVKDLAKVSVKEKLKMLRKESPEL---
LELIEDLKVKLTEVKDELE--PLLEL 250
++E ++ DL + E LK+L + PE L+ + LK +L

E+K++L+ P +L
10 Sbjct: 189 IEELVQQISPDLPR---
ELLKILEAKHPEFQLFLDELNQLKPQLNEIKEKLKTYPSSQL 245

Query: 251 VE 252
++

15 Sbjct: 246 LQ 247
Score = 57 (8.6 bits), Expect = 3.0e-01, Sum P(2) = 2.6e-01
Identities = 13/58 (22%), Positives = 26/58 (44%)

20 Query: 367 DSDFDKAKLKYYKEIEDRQKLKRK--
KEENSTEEQALEDAQNAKRAITYQIAKNRGLT 422
D + +++ L YY+ ++ + K+ +K KE + E + I
RG+T
Sbjct: 472
DREVEDQDDLDYYESLDKKSMAKKLRKENHDLERDLIRASRHPELIELGEGDKRGIT 529

25 Score = 42 (6.3 bits), Expect = 5.2e-09, Sum P(2) = 5.2e-09
Identities = 13/51 (25%), Positives = 29/51 (56%)

30 Query: 199 AETRVVKDLAKVSVKEKLKMLRKESPE--
LLELIEDLKVKLTEVKDELEPLLE 249
+ET + D+++ + LK ++++S + EL++ + L + EL
+LE
Sbjct: 160 SETDAIDDISQWADNSDLKSIKQDSSAAAIEELVQQISPDLP--
RTELLKILE 210

35 Score = 39 (5.9 bits), Expect = 1.1e-08, Sum P(2) = 1.1e-08
Identities = 8/18 (44%), Positives = 11/18 (61%)

40 Query: 43 YYQDQVDDFHEARSRAAL 60
+Y +Q+D RSRA L
Sbjct: 402 FYANQIDQKAAKRSRAVL 419

45 Pedant information for DKFZphamy2_2f22, frame 3

Report for DKFZphamy2_2f22.3

50 [LENGTH] 479
[MW] 54558.00
[pI] 5.50
[HOMOL] TREMBL:SPBC3B8_9 gene: "SPBC3B8.09"; product:
"hypothetical protein"; S.pombe chromosome II cosmid c3B8. 1e-10
55 [FUNCAT] 04.05.01.04 transcriptional control [S. cerevisiae,
YDL153c] 1e-08
[BLOCKS] PRO0528
[BLOCKS] BL00360C Ribosomal protein S9 proteins

[[BLOCKS]] BLO0964A Syndecans proteins
 [[BLOCKS]] PRO0624G
 [[BLOCKS]] PRO0828H
 [[BLOCKS]] BLO0824B Elongation factor 1 beta/beta'/delta chain

5 proteins

[[KW]] All_Alpha
 [[KW]] LOW_COMPLEXITY 24.63 %
 [[KW]] COILED_COIL 7.10 %

10

SEQ MVGRSRRRGAAKWAAVRAKAGPTLTDENGDDLGLPPSPGDTSYQDQVDDFHEARSRAAL
 SEGxx.....
 PRD cccccchhhhhhhhhhhhhhhccccccccccccccccccccccccccccccccchhhhhhhhhhhhhhh
 COILS

15

SEQ AKGWNEVQSGDEEDGEEEEEEVLALDMDDEDEDGGNAGEEEEEENADDDGGSSVQSEAE
 SEGxx.....
 PRD hhccccccccccccchhhhhhhhhhhhhhhccccccccchhhhhhhhhhhccccccchhhhhh
 COILS

20

SEQ ASVDPSLSWGQRKKLYD TDYGSKSRRGRQSQEAEEEEEEEEEAQIIQRRRLAQALQEDD
 SEGxx.....
 PRD hccccccccccccceeeccccccccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcc
 COILS

25

SEQ FGVAWVEAFAPVPQVDEAETRVVKDLAKVSVKEKLKMLRKESPELLELIEDLKVKLTEV
 SEGcc.....
 PRD chhhhhhhhhhhccccchhh
 COILS

30

SEQ KDELEPLLELVEQGIIPPGKGSQYLRTKYNLNLNYCSNISFYILKARRVPAHGHPVIER
 SEGcc.....
 PRD hhhhhhhhhhhhhhhhhccccchhh
 COILS

35

SEQ LVTYRNLINKLSVVDQKLSSEIRHLLTLKDDAVKKELIPKAKSTKPKPKSVSKTSAAACA
 SEGxx.....
 PRD hhh
 COILS

40

SEQ VTDLSDDSDFDEKAKLKYYKEIEDRQKLKRKKEENSTEEQALEDAQNAKRAITYQIAKNRG
 SEGcc.....
 PRD hhhhhccccchhh
 COILS

45

SEQ LTPRRKKIDRNPRVKHREKFRRRAKIRRRGQVREVRKEEQRYSGELSGIRAGVKKSIKLG
 SEGxx.....
 PRD cccccccccccccchhh
 COILS

50

55

(No Prosite data available for DKFZphamy2_2f22.3)

(No Pfam data available for DKFZphamy2_2f22.3)

5

DKFZphamy2_2g12

5 group: nucleic acid management

DKFZphamy2_2g12 encodes a novel 191 amino acid protein with similarity to NVL-2 of *Rattus norvegicus*.

10 The novel protein contains 3 EF-hand calcium-binding domains. The related human VILIP Ca-dependent protein specifically binds the 3'-untranslated region of the neurotrophin receptor, *trkB*, an mRNA localized to hippocampal dendrites in an activity-dependent manner. The new protein exhibits elevated expression in brain
15 and testis.

The new protein can find application in studying the expression profile of brain-specific genes and as a new marker for neuronal cells.

20

strong similarity to NVL-2 (*Rattus norvegicus*)

Comment for P35332:

25 FUNCTION: MAY BE INVOLVED IN THE CALCIUM-DEPENDENT REGULATION OF RHODOPSIN PHOSPHORYLATION.
TISSUE SPECIFICITY: NEURON-SPECIFIC IN THE CENTRAL AND PERIPHERAL NERVOUS SYSTEM.
MISCELLANEOUS: PROBABLY BINDS TWO OR THREE CALCIUM IONS (BY
30 SIMILARITY)
SIMILARITY: TO OTHER EF-HAND CALCIUM BINDING PROTEINS, BELONGS TO THE RECOVERIN SUBFAMILY.

35 Sequenced by MediGenomix

Locus: /chromosome="1"

Insert length: 4285 bp

40 Poly A stretch at pos. 4258, polyadenylation signal at pos. 4247

1 GCGGGCTCCG GCGCAGACCT TGGAGAGCAC AGCTGCCGGC CCGCGAGCCA
51 GCCTCGGTTT CCGCGGCCCG CCGAGGCTCG GAGCCATCCA GCGACCCGGC
45 101 GACCGGCCCTC AGGCCCCGCC ATGGGGGAAGA CCAACAGCAA GCTGGCCCCC
151 GAGGTGCTGG AGGACCTTGT TCAGAACACT GAGTTCAGCG AGCAGGAGCT
201 GAAGCAGTGG TACAAGGGCT TCCTGAAGGA CTGCCCCAGC GGCATCCTCA
251 ACCTGGAGGA GTTTCAGCAG CTCTACATCA AGTTCTTCCC CTACGGCGAC
301 GCCTCCAAGT TCGCGCAGCA CGCTTTCCGC ACCTTCGACA AGAACGGCGA
50 351 CGGCACCATC GACTTCCGGG AGTTCATCTG CGCCCTGTCT GTACCTTCCC
401 GCGGCAGCTT CGAGCAGAAG CTCAACTGGG CCTTTGAGAT GTACGACCTG
451 GACGGCGACG GCGGAATCAC GCGCCTGGAG ATGCTGGAGA TCATCGAGGC
501 AATCTACAAG ATGGTGGGCA CCGTGATCAT GATGCGCATG AACCAGGACG
55 551 GGCTCACGCC CCAGCAGCGT GTGGACAAGA TCTTCAAGAA GATGGACCAG
601 GATAAGGACG ACCAGATTAC ATTGGAGGAG TTCAAGGAGG CAGCCAAGAG
651 TGACCCATCC ATTGTGTTGC TGCTGCAGTG TGACATGCAG AAGTAGAAGC
701 TGGTGAGGGG CAGGGTCCCT GGCCAGAAGG GGCATGGCCA CCTCCCAACC
751 TGATGACCTC TCTGGCTGGC CTCCAGGAG GAGGGACACT CCAGCCCCC

	801	TCTCTGGCCC	ACCCAGTCCT	CTGCCCCAAGC	CCTTCCTCCC	CTCCATCAAG
	851	ATCTTTTGAGG	GACCACCTCA	CCCTGCAAAA	GAGACAGGTC	CTCCAGTACC
	901	CTGTCTTCTA	GCCCCACCTC	CCACTTGGCC	AGAACCAATG	TCCATTGGGC
	951	ATAGGGGAGT	TGGCTTTTGC	CCCAGGAGGT	GAGGTTAAGG	AGTTGGGGGC
5	1001	CTGGGGTTCT	GGTTAGGAAT	TCTCTTGATC	CTGGGATTAT	GCTTTATAGG
	1051	ATGTGGTCCC	ACAGGCCTGT	CACAGGGCCA	AATTGGGTCT	GTCCATTCTC
	1101	GAGGCTCCAG	ATCCCATAAA	GGGGGTCTCT	TCCCCATCCC	TTCTACTCTA
	1151	CCTGGCCCTT	CCAGCCCCAG	CCTTTGGAGC	GTTCAATTCAG	TCCTTTCTTC
	1201	AGCTAATGAT	TACTGAGCAC	CTGTTTGGTG	CTAAGGATAT	GGTCATTTAC
10	1251	AAGACACATC	TTGTGCCCTC	TGGAAGCTCA	TAGGGTTGTG	AGGCAAACTT
	1301	CCAGCCGTCA	GGGTCTCAGC	TAAGCAGAAG	GTGCTGGAAG	GCTGGTTAGT
	1351	CTGGGAGGAG	CTATTTTCAT	TCCAGCTCA	GCTCCACACA	AAGCTGCAGA
	1401	AGGACGAAAT	GAAAAGCATT	TGGAAGTTTA	GGAGCCACGT	GAGTGAAAGT
	1451	TTTAAGAAAA	ATGAAATTTA	TGTCATACTT	ATTTTTTTAG	TACCTTTTAA
15	1501	AGGAGCTACA	GTCATTTTAT	TATTTTCAGGA	GGTTAAAATA	TACTCTATAT
	1551	TACTTGTTTT	ATTATAAAAT	GATTAAATGA	ATAGAGAAAA	TATTAATTTT
	1601	CAAGGGGAAA	AAACCTGAGA	AGAAAGGGAG	AAAAGACCAT	GAAATTTACC
	1651	AGATAACACT	TTTTAAGACT	AAGTCCTGAG	CTGCCACTCT	CAGCAGTTTT
	1701	TGCTGCTTCA	GCTCTTCCTT	TTTATTACCT	TTTTCAATTC	AACAAGCAAC
20	1751	TTTCTGCTAC	ATACTTACTC	CGGTTGGGTG	CTGACTTCAG	GGACAGGAAA
	1801	AAGCAAGGTT	TGCAAAAGAGT	GAAACTAGTG	TATATTCCGT	ATCTTGGTAG
	1851	TTCGTTTCTG	GATTGGGTTT	AGTTTCAGAA	CTGGACTTGT	TCCTTCACTG
	1901	CCACAGAATC	AGAAAGAGCT	AGAAGAAAAG	GCTCACCTGG	CCACTGTTTA
	1951	GGCACCCAGA	CATAATTTAT	GGACGAAATG	CCTAAAAATG	TGCCAGGCAT
25	2001	GCTCTGTTTT	AGAGGCTTTT	TCTAACCCCA	AATCTTAGAT	CTGCCAGGTA
	2051	GTTCAACATC	TTCCAAGTGT	GCTGGTTCTG	CTTTCCAATG	CCTGCTTCCC
	2101	AATTTTGGAT	CCATGAGCTA	TACAGCTGCA	TGCTTTGACT	GCCGGAAAAA
	2151	TTAATCTTGC	TTCTTCATCA	GGTCTTTCTC	CTGTACTTGT	GATCAGAAAT
	2201	TACCTTTGAC	GTGCAGTGAC	AGTTGATTTC	CTCTGAACT	GCCGGTGAAA
30	2251	ACAGTCTAGT	ACACAGGTGC	TGTCAGCCCA	GGGTGGGAGC	AGGAAATGAT
	2301	TGCTGAGCCC	GGGGCAGGGG	AATTGCATCT	GCAGGAAAGA	GATGCAGCAT
	2351	GCTCCTCACT	CCTGAGTGCC	CACCTGTCCT	GCTTCTCTGC	AGGTGAAAAC
	2401	TCTGGGGGAT	GCTGATCAAT	AGAGCTTGGT	CCCAAGCTCT	ACTGGGCCCT
	2451	TGGAGGTAGC	AAGGCCACTG	GGTTGCTATC	CTCTTGATGG	GGATAGCAAC
35	2501	CACTGGTTTT	CAACCACTGG	GTTGCTATCC	TTTTGCTATC	CTCTTGCTCA
	2551	TGACCAGCCA	TATGGTGAGG	CTGGGGAGTT	CACATCCTCA	GGCAGGAACT
	2601	AGCAGTTGTT	TATCCAGCAA	TGCCCAAGG	ATGTTGCATT	GCTCCCAGGA
	2651	GCTGGCTATT	AGGTATGTCT	TGTGCGGTCA	GTCAGCATCA	CAGACACATA
	2701	GATGCTCACC	AGCCTGGCTT	AGCTGGGACC	TAAATCTTCT	GGTGAAAAGC
40	2751	TTTTCACTAA	GTGAGGTTCC	TTCCCTGCAA	ATGCTGAATC	TAGCCTAATT
	2801	CGCAACCACA	CAGAATTTCA	TGGCTTTCAA	AGGCTTGCCA	TGTGCCCCAT
	2851	CTCATTCTAT	ACTCACATCC	CATGGAGGTG	AGGATTTTCA	CTTCTTTTCT
	2901	CTAGACTTGG	AAGCTGAGAT	TCAGAGAGGA	AGCATCCCTT	GTGCAAGATC
	2951	ACATAGTCAG	GAGGTGACAC	AGGGCTAAGA	CTTGAACCAA	GGCTCTAAGA
45	3001	GGATTTCTTC	TTTTTCAGAGT	CTCTTCCCTG	TCCATTTCTG	TGACTAAGCT
	3051	GTGCAGAGGT	TGACAGCAGG	GCAAGTTACA	TTGATATTCA	TTCTTTATAG
	3101	GCTTCCTGCT	AAAAAGCTTC	TGAGATTGTG	GTCTTCCAAA	AAAAATAGGA
	3151	GCTTGGTTGA	AGTCCCCACA	TTTTCAAGCA	CTCAGTGTTT	TGCCTCTGGC
	3201	AGCTGTGCTA	ACAGCTCAGT	GCTGTCCTGG	GAGTCCTCTG	ACTCAGAAC
50	3251	CTCGAAGCAT	CCTGCATTGT	CTTTACCCAC	CATCATCGTC	ACTAAGAGAA
	3301	ACATGCCTAC	CCATGAAGGC	GTGTTTGATT	ACTCCAGGCT	TCTGGACACA
	3351	CATACCCATG	GGTGATTTTT	GCTCCTCAGG	CCCAATATTC	TCAGACAGCC
	3401	CAGCAGTGTG	AACACACAAT	GCCAGGCCAG	GAAGTGGGAC	CACCATCTTG
	3451	CTGATGGAAG	GAACAACAGG	TGGCCCAGGA	CATGCTCCTG	CATACTCCTG
55	3501	GGTGTCCCA	GGACTGTGTG	CTCAGGAGCA	CTGTGGTAGA	GCCTGGCCC
	3551	TGCCTTGAGA	AGAGACACAG	GTCTCCCGTC	CCTGCACCAG	CTGAGAGAGA
	3601	CTTGCCACAA	AGCACAAAGC	TGGCAGAGAT	TTATGTATGA	CTTGACACAGA
	3651	CACAAAAATA	TACAGACAAT	CAAAACATTG	ATATATTCAA	ACTCTCCTTT

```

3701 AAATTCCAAT CTTATTGCAA CAACTCTGTG AATTGCAAGG TCCCAGAATC
3751 TGCCTTCTCA CATACTCTAC CCTCATTCAT CCTTTTGGGC TAATTGATGA
3801 GCATCTTATT TCTTATCTCT AAAAATTATC AGCAAAGGCT ACTTCAGATG
3851 GCCACTTTAG TCCTTTCAGC TGTAAGTCAGG ATTATTTAAC TTACCTGTAT
5 3901 ATCAAAAAGTG AAGAAAAAGT TAGTTCATAA GTAAAGGCAC TAAATCCTTT
3951 CCTGACAATG GCAGAGTCTC TAGAGGTAGA AATTTGCCTT GCTGCAGAGA
4001 GAGAAGGAAT GGCCTGGGAT GGGGGAAAAG AAAGAAAGAG AAGAAGAGAA
4051 GAAGCTGGGG TCTCCAGGCA GGGTAGTAAG CTGACACTAA ATATTTTTTA
4101 CACAAAAATG TATTGAAGCA ACAAATATTT CCTGAAGATC CACCCTGGGT
10 4151 GAGGCTTTGA GCTGACTTTA GAGATCACTG TGGGGTCAAG AATGTCTTAC
4201 ATGTTTTATT CATATTCTT GAAAAAAGAA ATAATTCAAA CCTTGGAATT
4251 AAAAAGTCAG AAAAACAAAA AAAAAAAAAA AAAAA

```

15 BLAST Results

No BLAST result

20 Medline entries

93367470:

25 Kajimoto Y, Shirai Y, Mukai H, Kuno T, Tanaka C.; Molecular
cloning of
two additional members of the neural
visinin-like Ca(2+)-binding protein gene family. J Neurochem 1993
Sep;61(3):1091-6

96079121:

30 Polymeropoulos M.H., Ide S., Soares M.B., Lennon G.G.; Sequence
characterization and genetic mapping of the human VSNL1 gene, a
homologue of the rat visinin-like peptide RNP1. Genomics
35 29(1):273-275(1995).

40 Peptide information for frame 1

ORF from 121 bp to 693 bp; peptide length: 191

Category: strong similarity to known protein

45 Classification: Protein management

Prosite motifs: EF_HAND (73-85)

EF_HAND (109-121)

EF_HAND (159-171)

```

50 1 MGKTNSKLAP EVLEDLVQNT EFSEQELKQW YKGFLKDCPS GILNLEEFQQ
51 LYIKFFPYGD ASKFAQHAFR TFDKNGDGTI DFREFICALS VTSRGSFEQK
101 LNWAFEMYDL DGDGRITRLE MLEIIIEAIYK MVGTVIMMRM NQDGLTPQQK
55 151 VDKIFKKMDQ DKDDQITLEE FKEAAKSDPS IVLLLQCDMQ K

```

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_2g12, frame 1

No Alert BLASTP hits found

Pedant information for DKFZphamy2_2g12, frame 1

Report for DKFZphamy2_2g12.1

```

15  [LENGTH] 231
    [MW] 26277.92
    [pI] 5.26
    [HOMOL] PIR:JH0815 neural visinin-like Ca2+-binding
    protein-type 2 - rat 1e-107
    [FUNCAT] 98 classification not yet clear-cut [S. cerevisiae,
20  YDR373w] 3e-52
    [FUNCAT] 03.01 cell growth [S. cerevisiae, YKL190w] 3e-18
    [FUNCAT] 03.07 pheromone response, mating-type determination,
    sex-specific proteins [S. cerevisiae, YKL190w] 3e-18
    [FUNCAT] 13.04 homeostasis of other ions [S. cerevisiae,
25  YKL190w] 3e-18
    [FUNCAT] 04.05.01.04 transcriptional control [S. cerevisiae,
    YKL190w] 3e-18
    [FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae,
    YKL190w] 3e-18
30  [FUNCAT] 11.01 stress response [S. cerevisiae, YGR100w] 7e-04
    [BLOCKS] BL00303B S-100/ICaBP type calcium binding protein
    [BLOCKS] BL00018
    [BLOCKS] PR00450G
    [BLOCKS] PR00450F
35  [BLOCKS] PR00450E
    [BLOCKS] PR00450D
    [BLOCKS] PR00450C
    [BLOCKS] PR00450B
    [BLOCKS] PR00450A
40  [SCOP] d1osa_ 1.37.1.5.13 Calmodulin [(Paramecium
    tetraurelia) 8e-25
    [SCOP] d1rec_ 1.37.1.5.21 Recoverin [bovine (Bos
    taurus) 1e-72
    [SCOP] d1a4pa_ 1.37.1.2.5 Calcyclin (S100) [Human (Homo
45  sapiens), P1 7e-05
    [SCOP] d1rro_ 1.37.1.4.1 Oncomodulin [rat (Rattus
    norvegicus) 2e-17
    [SCOP] d1syms_ 1.37.1.2.2 Calcyclin (S100) [rat (Rattus
    norvegicus) 9e-14
50  [SCOP] d4icb_ 1.37.1.1.1 Calbindin D9K [bovine (Bos
    taurus) 2e-18
    [SCOP] d1auib_ 1.37.1.5.19 Calcineurin regulatory subunit
    (B-chain 1e-45
    [PIRKW] blocked amino end 1e-99
55  [PIRKW] phosphotransferase 3e-08
    [PIRKW] duplication 7e-17
    [PIRKW] tandem repeat 7e-06
    [PIRKW] heterodimer 7e-17

```

[[PIRKW]] heart 7e-06
 [[PIRKW]] serine/threonine-specific protein kinase 7e-06
 [[PIRKW]] acetylated amino end 7e-06
 [[PIRKW]] ATP 7e-06
 5 [[PIRKW]] skeletal muscle 7e-06
 [[PIRKW]] signal transduction 4e-69
 [[PIRKW]] protein kinase 3e-08
 [[PIRKW]] calcium binding 1e-99
 [[PIRKW]] alternative splicing 1e-13
 10 [[PIRKW]] lipoprotein 1e-99
 [[PIRKW]] cardiac muscle 7e-06
 [[PIRKW]] muscle 7e-06
 [[PIRKW]] myristylation 1e-99
 [[PIRKW]] EF hand 1e-99
 15 [[PIRKW]] retina 1e-46
 [[SUPFAM]] calcium-dependent protein kinase 3e-08
 [[SUPFAM]] unassigned calmodulin-related proteins 2e-34
 [[SUPFAM]] protein kinase homology 3e-08
 [[SUPFAM]] calmodulin 1e-99
 20 [[SUPFAM]] calmodulin repeat homology 1e-99
 [[PROSITE]] EF_HAND 3
 [[PFAM]] EF hand
 [[KW]] All_Alpha
 [[KW]] 3D

25

SEQ GGS GADLGEHSCR PASQPRFPRPAEARSHPATRRPASGPAMGKTNSKLAPVLEDLVQNT
 1rec-

30

.....HHHHHHHHHTTTT

SEQ EFSEQLKQWYKGF LKDCPSGILNLEEFQQLYIKFFPYGDASKFAQHAFTFDKNGDGTI
 1rec- CCCHHHHHHHHHHHHHHTTTTTEEHHHHHHHHHHHHTTTTCHHHHHHHHHHHHH---
 --CEE

35

SEQ DFREFICALSVTSRGSFEQKLNWAFEMYDL DGDGRITRLEMLEIIEAIYKMVGTVIMMRM
 1rec- EHHHHHHHHHHHHHCCCGGGHHHHHHHHHTTTTCCCEEHHHHHHHHHHHHHCCTTTTGGGC

40

SEQ NQDGLTPQQRVDKIFKKMDQDKDDQITL EEFKEAAKS DPSIVLLLQCDMOK
 1rec- TTTTCHHHHHHHHHHHHCCTTTTEECHHHHHHHHHHHCHHHHHHHCCCHH

Prosite for DKFZphamy2_2g12.1

45

PS00018	113->126	EF_HAND	PD0C00018
PS00018	149->162	EF_HAND	PD0C00018
PS00018	199->212	EF_HAND	PD0C00018

50

Pfam for DKFZphamy2_2g12.1

55

HMM_NAME EF hand

HMM *EIqEMFrMMDkDGDGyIDFEEFmeMMkem*
 Q +FR +DK+GDG+IDF EF+ +++

Query 104 FAQHAFRTFDKNGDGTIDFREFICALSVT 132

27.15 140 168 1 29 dkfzphamy2_2g12.1 strong
similarity to NVL-2 (Rattus norvegicus)

5 Alignment to HMM consensus:
Query *EIqEMFrMMDkDGDGyIDFEEFmeMMkem*
++++F+M+D DGDG+I+ E++E++ ++
dkfzphamy2 140 KLNWAFEMYDL DGDGRITRLEMLEIIEAI 168

10 Query 218 1 29 dkfzphamy2_2g12.1 strong
similarity to NVL-2 (Rattus norvegicus)
Alignment to HMM consensus:
HMM *EIqEMFrMMDkDGDGyIDFEEFmeMMkem*
++++F++MD+D+D +I+ EEF+E+ K+

15 Query 190 RVDKIFKKMDQDKDDQITLEEFKEAAKSD 218

DKFZphamy2_2117

5 group: amygdala derived

DKFZphamy2_2117 encodes a novel 462 amino acid protein without similarity to known proteins.

10 Most ESTs are derived from brain and pancreas.
No informative BLAST results; No predictive prosite, pfam or SCOP motife.

15 The new protein can find application in studying the expression profile of amygdala-specific genes.

unknown protein

20 perhaps complete cds.

Sequenced by MediGenomix

Locus: unknown

25

Insert length: 3473 bp

Poly A stretch at pos. 3454, polyadenylation signal at pos. 3436

30 1 GATATCCCAA TCTTTGGACT GCATCCTGGT TGCCTCTACT GTGGTCACCT
51 TTGGGAAGAA ATGTCTTCTG TAAAAAGAAG TCTGAAGCAA GAAATAGTTA
101 CTCAGTTTCA CTGTTCAAGT GCTGAAGGAG ATATTGCCAA GTTAACAGGA
151 ATACTCAGTC ATTCTCCATC TCTTCTCAAT GAAACTTCTG AAAATGGCTG
201 GACTGCTTTA ATGTATGCGG CAAGGAATGG GCACCCAGAG ATAGTCCAAT
35 251 TTCTGCTTGA GAAAGGGTGT GACAGATCAA TTGTCAATAA ATCAAGGCAG
301 ACTGCACTGG ATATTGCTGT ATTTTGGGGT TATAAGCATA TAGCTAATTT
351 ACTAGCTACT GCTAAAGGTG GGAAGAAAGCC TTGGTTCCCTA ACGAATGAAG
401 TGGAAGAATG TGAAAATTAT TTTAGCAAAA CACTACTGGA CCGGAAAAGT
451 GAAAAGAGGA ATAATTCTGA CTGGCTGCTA GCTAAAGAAA GCCATCCAGC
40 501 CACAGTTTTT ATTCTTTTCT CAGATTTAAA TCCCTTG GTT ACTCTAGGTG
551 GCAATAAAGA AAGTTTCCAA CAGCCAGAAG TTAGGCTTTG TCAGCTGAAC
601 TACACAGATA TAAAGGATTA TTTGGCCCAAG CCTGAGAAGA TCACCTTGAT
651 TTTTCTTGGA GTAGAAGTTG AAATAAAAGA CAAACTACTT AATTATGCTG
701 GTGAAGTCCC GAGAGAGGAG GAAGATGGAT TGGTTGCCTG GTTTGCTCTA
45 751 GGTATAGATC CTATTGCTGC TGAAGAATTC AAGCAAAGAC ATGAAAATTG
801 TTACTTTCTT CATCCTCCTA TGCCAGCCCT TCTGCAATTG AAAGAAAAAG
851 AAGCTGGGGT TGTAAGCTCA GCAAGATCTG TTCTTGCTG GCACAGTCGA
901 TACAAGTTTT GCCCAACCTG TGGAAATGCA ACTAAAATTG AAGAAGGTGG
951 CTATAAGAGA CTATGTTTAA AAGAAGACTG TCCTAGTCTC AATGGCGTCC
50 1001 ATAATACCTC ATACCCAAGA GTTGATCCAG TAGTAATCAT GCAAGTTATT
1051 CATCCAGATG GGACCAAATG CCTTTTAGGC AGGCAGAAAA GATTTCCCCC
1101 AGGCATGTTT ACTTGCTTTG CTGGATTTAT TGAGCCTGGA GAGACAATAG
1151 AAGATGCTGT TAGGAGAGAA GTAGAAGAGG AAAGTGGAGT CAAAGTTGGC
1201 CATGTTCAAGT ATGTTGCTTG TCAACCATGG CCAATGCCTT CCTCCTTAAT
55 1251 GATTGGTTGC TTAGCTCTAG CAGTGTCTAC AGAAATTAAA GTTGACAAGA
1301 ATGAAATAGA GGATGCCCGC TGGTTCACTA GAGAACAGGT CCTGGATGTT
1351 CTGACCAAAAG GGAAGCAGCA GGCATTCTTT GTGCCACCAA GCCGAGCTAT
1401 TGCACATCAA TTAATCAAAC ACTGGATTAG AATAAATCCT AATCTCTAAA

1451 TCTAAGAACT AAGCTTTGAG TATTATTTAA TAATTTCTAA TAACACTCAT
1501 TCCTCAAGTG ATATTAGAGA TTATTCAGTA CTCTTGAGAG TGTCAACAACA
1551 CAAAATACGA TGTGGGGTTT TCGAAATATT TTCAAAGTGT TCTGTCTTAA
1601 TCACAAATTC ATATTTTAC ACATTTTAC AATATTGCCT CAGATTATGT
5 1651 TAAATTTGGG TCAGTCTTCT CTGAACTTT TCTCTCTCGG TTTCTTTTCT
1701 TCCTTCACAG TTTTATCTCA CAAAACCAT TTTCTAATAA GAGACATCAT
1751 GTTGGAAAGA TGTGTAGAA ATGTGCATAA ATTTCAAGTGC CTCTTGTAAG
1801 CATTAAACTG ATGATGAAGA AAGTTCCTGA TTTGAGAAAT GAATCAAAGT
1851 AATTTTAATG AATTTTATAGC TTGTATTAGC TTGAGTTAGC TGGCATTGAT
10 1901 TTTTATAGTCC TTTTGTACC TTTAAGTTGT CAATATATGG TTTTGTTC
1951 TCTCCCCATT GTAGTCCCAC TTGCTCTTTC CTGGGGGTTC CATTGTTCTA
2001 GCAGTGGAGG GTTACAGTG TCGCCACTCG TCTAATTGTA CCAGTGTTAA
2051 GAATTTTCTA ATTTAATAAT TTAATAGTGA TCTCAATACC ACACCTCAT
2101 GGAAGGAGAA AAGCATACTA TTATATCTGG GACCTCTCTT TTAGACCTAA
15 2151 AATTAATTAA CATATCTACT TATATGTTAC TTATACCTAA AGCTGTTATT
2201 AAGACAAACC AAGATTCTCT GCTTTTGCAC TGAAATTAAT CTTGAAAGGA
2251 ATTCTCCTCA AAGGTCGGAT ATTAATAAAG TCCCAGGCAG ATTTACATAT
2301 TTAATTTAAA ACATTGGCTT TATTTTCAAT TGTGATGAGT GATGTATCTG
2351 TGTTAAACAAA AAATTGTATA ATCATTACCA ATACTATTTA TTATGCTCAA
20 2401 ATATATCTTG GCTTTGACCT TATTTCAACA CATTCTAAGA AGCCTTGACA
2451 AAGTAAGTAT ATTTTAGAGC TGAATCAGTA AGATTCTAGA GAAAGCAAAA
2501 CATAGTAGTT CACAATTTTG CAACATAGAA AGTCACATTT TGAAAGGCTA
2551 TTTTGAAATT GATTTAATAG CTATTATAGT TTATGAATAT CAAAATTTGT
2601 ATAATTTGCA TCTTTACTAA TGTATGCTAG AGCTACAAGA GACCTTAAGG
25 2651 ATAATATATG AAATTAGCTT TCCTTATTTT ATAGATAAGG AAAAAGAAAT
2701 TGTGAAAGGT GAATTTACCT AATTAGTGAA AGTTACATAA CTAATTACAA
2751 CAGTCTGTAC TATATAATGC AGAGGACGAT TCTCCCTGTA AAAGGAACATA
2801 GAAGCTATTA CTAATAATAT ATATAGACAA AATTAATAAG AGGAATGATA
2851 AGAATAAATT TAATTTACCA AATATTGTTA ATTAATAATTT TAGATACTTA
30 2901 ACATTTATTT AACTTAAATA AAAGATAACT GTCAGATAAA ACTTTATTTT
2951 ACTAATGAGC AGTGATTTTC TTAGGAATTG ATGAAGGCTT ATTGGTATCA
3001 AGAATTTAAA CCAATTTAAA ACTGACAGAG GACATTTAGA TACATAATAA
3051 AATTCGAGCT ACATAAGTAT ATGGAAAATA ATGTACCTTG ATTATTATGA
3101 AATAGAGCAT CTTGAAATTC AGTTTTACTC TAAATGTACT TTTAATACTT
35 3151 GCAGATTCTA AGATTACATT GTGAAATTCC AGGTTTTTCAT AATGTTAAAA
3201 TAGGAAAGTA GAATATAAAG TATCAACAAG TGTAAGTTATA CATTTTGT
3251 TGGATATTTA ATCCTTACTT GGGAAAAAAT CAGCATCTAG GTAAATTATT
3301 ATTTTAATAA GAACTCTTAA ATTGCCAACC TCTGAGAGGT GAAAAGCTAT
3351 GTAAATAGAA GGAATGGCCA GTTCAAAAAG ATAGTAGAAG TGATAGTGCC
40 3401 GTGAATGTAT TCTACTGGAA ATGAATGTAA TAATACATTA AATTTTAA
3451 ATCGAAAAAA AAAAAAAAAA AAA

BLAST Results

No BLAST result

Medline entries

No Medline entry

Peptide information for frame 1

ORF from 61 bp to 1446 bp; peptide length: 462

Category: putative protein

Classification: unclassified

5 Prosite motifs: MUTT (355-374)

```

10 1 MSSVKRSLKQ EIVTQFHCSA AEGDIAKLTG ILSHSPSLLN ETSENGWTAL
    51 MYAARNHGHE IVQFLLEKGC DRSIVNKSQ TALDIAVFWG YKHIANLLAT
    101 AKGGKKPWFL TNEVEECENY FSKTLLDRKS EKRNSDWLL AKESHPATVF
    151 ILFSDLNPLV TLGGNKESFQ QPEVRLCQLN YTDIKDYLAQ PEKITLIFLG
    201 VELEIKDKLL NYAGEVPREE EDGLVAWFAL GIDPIAAEEF KQRHENCYFL
    251 HPPMPALLQL KEKEAGVVAQ ARSVLAWHSR YKFCPTCGNA TKIEEGGYKR
    301 LCLKEDCPSL NGVHNTSYPR VDPVVMQVI HPDGTKCLLG RQKRFPFGMF
    15 351 TCLAGFIEPG ETIEDAVRRE VEEESGVKVG HVQYVACQPW PMPSSLMIGC
    401 LALAVSTEIK VDKNEIEDAR WFTREQVLDV LTKGKQQAFF VPPSRAIAHQ
    451 LIKHWIRINP NL

```

20

BLASTP hits

No BLASTP hits available

25

Alert BLASTP hits for DKFZphamy2_2117, frame 1

No Alert BLASTP hits found

30

Pedant information for DKFZphamy2_2117, frame 1

Report for DKFZphamy2_2117.1

```

35 [LENGTH] 462
    [MW] 52076.25
    [pI] 6.38
    [HOMOL] TREMBL:SPBC1778_3 gene: "SPBC1778.03c"; product:
    "conserved hypothetical protein"; S.pombe chromosome II cosmid
40 c1778. 1e-45
    [FUNCAT] 99 unclassified proteins [S. cerevisiae, YGL067w]
    4e-34
    [FUNCAT] r general function prediction [H. influenzae,
    H10432 pyrophosphohydrolase] 4e-24
45 [FUNCAT] 1 genome replication, transcription, recombination and
    repair [M. jannaschii, MJ1149 nucleotide pyrophosphohydrolase]
    1e-04
    [BLOCKS] BL00219F Anion exchangers family proteins
    [BLOCKS] BL01293B
50 [BLOCKS] DM01909
    [BLOCKS] PF00023A
    [BLOCKS] BL00893 mutT domain proteins
    [SCOP] dlawcb_ 1.91.3.1.2 GA binding protein (GABP) alpha
    GA bindini 2e-35
55 [SUPFAM] hypothetical protein H10432 1e-22
    [PROSITE] MUTT 1
    [PFAM] Bacterial mutT protein
    [PFAM] Ank repeat

```

[[KW]] Irregular
[[KW]] 3D

```

5  SEQ  MSSVKRSLKQEIIVTQFHCSAAEGDIAKLTGILSHSPSLLNETSENGWTALMYAARNGHPE
lawcB  .CCCTTTTCTTTCCHHHHHHHTTHHHHHHHHCCCTT-
TTEETTTEHHHHHHHHHCCHH

10  SEQ  IVQFLLEKGCDRSIVNKSRTALDIAVFWGYKHIANLLATAKGGKKPWFLTNEVEECENY
lawcB  HHHHHHHHCCTTTTTCBTTTBCHHHHHHHHCCHHHHHH.....

15  SEQ  FSKTLLDRKSEKRNSDWLLAKESH PATVFILFSDLNPLVTLGGNKESFQQPEVRLCQLN
lawcB  .....

20  SEQ  YTDIKDYLAQPEKITLIFLGVELEIKDKLLNYAGEVPREEEDGLVAWFALGIDPIAAEEF
lawcB  .....

25  SEQ  KQRHENCYFLHPPMPALLQLKEKEAGVVAQARSVLAWHSRYKFCPTCGNATKIEEGGYKR
lawcB  .....

30  SEQ  LCLKEDCPSLNGVHNTSYPRVDPVVIMQVIHPDGTKCLLGRQKRFPFGMFTCLAGFIEPG
lawcB  .....

35  SEQ  ETIEDAVRREVEEESGVKVGHVQYVACQWPMPSSLMIGCLALAVSTEIKVDKNEIEDAR
lawcB  .....

    SEQ  WFTREQVLDVLTGKGQQAFFVPPSRAIAHQLIKHWIRINPNL
lawcB  .....

```

Prosites for DKFZphamy2_2117.1

40 PS00893 355->375 MUTT PD0C00695

Pfam for DKFZphamy2_2117.1

45

HMM_NAME Ank repeat

```

50  HMM          *GyTPLHIAARYNNvEMVr1LLQHGADIN*
      G+T+L++AAR+++ E+V++LL++G D
Query          46  GTALMYAARNGHPEIVQFLLEKGCDRS 73

```

55 HMM_NAME Bacterial mutT protein

```

HMM
*ILMiqRedppnHYdtHhgDWIFPGGkIEeGETPEQCarREIWEETGI*

```

L++++++ + +
++G+IE+GET+E+++RRE++EE+G+
Query 337 CLLGRQKRF--PPG----
MFTCLAGFIEPGETIEDAVRREVEEESGV 377

5

DKFZphamy2_2013

5 group: intracellular transport and trafficking

DKFZphamy2_2013 encodes a novel 590 amino acid protein with high similarity to murine synaptotagmin 3.

10 The novel protein contains two C2 domains. The C2 domain is thought to be involved in calcium-dependent phospholipid binding. Synaptotagmins are essential for Ca(2+)-regulated exocytosis of neurosecretory vesicles.

15 The new protein can find application in modulating/blocking synaptic activity.

similarity to synaptotagmin 3 (Mus musculus)

20 Sequenced by MediGenomix

Locus: unknown

25 Insert length: 2931 bp

Poly A stretch at pos. 2912, polyadenylation signal at pos. 2884

```

1  ACTCTATGTC TCCTCTCGTT GGATTGTGAC ACCGGGAGGT CAGGGAACTC
30  51  CAGGACCTTG TTCTCTGCTG GATTTCGCAGC AACCAGCACA GCACGTAGGG
    101  CGTAGTTGGT GCTGGATGGA TGTTTGTTGA ATGAATGAAT GATGAATGGC
    151  TGGCACCTTG TCTGCTCATC CCTAACTCCT GTTCCTTCAT CTGTGCAGCC
    201  CTAATCTTTG TTTCCTCATC TGTCCATCCC TTTATTTGTG CATCCTCATT
    251  CTTAGCCCCCT TCACTGCCCT TCTCCATCTC TTCCTCCTTG TTCATTTGTC
35  301  CCTGTTCTCT GTCTCTACT CCACTCATGC CCATCTCTGT CCCCTTGACT
    351  TACCCAGTCC CTGCTACTAT CTCCATCCCT AATTTCTGCC CTCTTGCTCTG
    401  TCTACTCCTA ATTCTTTTTC CTGTCCATC CCTAATACCT GTCACCTTGT
    451  CTTTCTTCCT CGAATCTCCA TCCCTAATCC ATCTGCCCTT AATCTCTGTC
    501  CCCTTTGCCC ATCTTCTCTT TTCTCGGTGT CTCTTTCCAC CCTTATCTCC
40  551  ACACCTGCCC ACCCTGCACT CCCATTCTGT TTCCCATCTG CACCCTTGCC
    601  CCATCCCTCC CACACACAGG ACCAGACGGC CACCATGTCA GGAGACTACG
    651  AGGATGACCT CTGCCGGCGG GCACTCATCC TGGTCTCGGA CCTCTGTGCG
    701  CGGGTCCGAG ATGCTGACAC CAACGACAGG TGCCAGGAGT TCAATGACCG
    751  AATCCGAGGC TATCCCCGGG GTCCAGATGC AGACATCTCC GTGAGCCTGC
45  801  TGTGGTTCAT CGTGACATTC TGTGGCATTG TCCTTCTGGG TGTCTCTCTC
    851  TTCGTGTCCT GGAAGTTGTG CTGGGTGCCC TGGCGGGACA AGGGAGGCTC
    901  GGCAGTGGGC GGTGGCCCCC TGCGCAAAGA CCTAGGCCCT GGTGTGCGGC
    951  TGGCAGGCCT GGTAGGCGGA GGCGGGCACC ACCTGGCGGC TGGCCTGGGT
50  1001  GGCCATCCTC TGCTGGGCGG CCCACACCAC CATGCCCATG CCGCCACCA
    1051  TCCACCCTTT GCTGAGCTGC TGGAGCCAGG CAGCCTGGGG GGTTCAGACA
    1101  CCCCTGAGCC CTCCTACTTG GACATGGACT CGTATCCAGA GGCTGCAGCA
    1151  GCAGCAGTGG CCGCTGGGGT CAAACCGAGC CAAACATCCC CTGAGCTGCC
    1201  CTCTGAGGGG GGAGCAGGCT CTGGGTGCTT CCTGCTGCCC CCCAGTGGTG
    1251  GGGGCTTGCC CAGTGCCCA GTCACATCAGC AGGTCAACA GCTGGCACCC
55  1301  ACTACCAGGT ACCCAGCCCT GCCCCGACCC CTCACCCAGC AGACTCTGAC
    1351  CTCCCAGCCG GACCCAGCA GTGAGGAGCG CCCACCTGCC CTGCCCTTAC
    1401  CCCTGCCTGG AGGCGAGGAA AAAGCCAAAC TCATTGGGCA GATTAAGCCA
    1451  GAGCTGTACC AGGGGACTGG CCCTGGTGGC CGGCGGAGCG GTGGGGGCCC

```

```

1501 AGGCTCTGGA GAGGCAGGCA CAGGGGACC CTGTGGCCGT ATCAGCTTCG
1551 CCCTGCGGTA CCTCTATGGC TCGGACCAGC TGGTGGTGAG GATCCTGCAG
1601 GCCCTGGACC TCCCTGCCAA GGA CTCCAAC GGCTTCTCAG ACCCTACGT
1651 CAAGATCTAC CTGCTGCCTG ACCGCAAGAA AAAGTTTCAG ACCAAGGTGC
5 1701 ACAGGAAGAC CCTGAACCCC GTCTTCAATG AGACGTTTCA ATTCTCGGTG
1751 CCCCTGGCCG AGCTGGCCCA ACGCAAAC TG CACTTCAGCG TCTATGACTT
1801 TGACCGCTTC TCGCGGCACG ACCTCATCGG CCAGGTGGTG CTGGACAACC
1851 TCCTGGAGCT GGCCGAGCAG CCCCCTGACC GCCCGCTCTG GAGGGACATC
1901 GTGGAGGGCG GCTCGGAAAA AGCAGATCTT GGGGAGCTCA ACTTCTCACT
10 1951 CTGCTACCTC CCCACGGCCG GGCGCCTCAC CGTGACCATC ATCAAAGCCT
2001 CTAACCTCAA AGCGATGGAC CTCCTGGCT TCTCAGACCC CTACGTGAAG
2051 GCCTCCCTGA TCAGCGAGGG GCGCGTCTG AAGAAGCGGA AAACCTCCAT
2101 CAAGAAGAAC ACGCTGAACC CCACCTATAA TGAGGCGCTG GTGTTGACG
2151 TGGCCCCCGA GAGCGTGGAG AACGTGGGGC TCAGCATCGC CGTGGTGGAC
15 2201 TACGACTGCA TCGGGCACA CGAGGTGATC GCGGTGTGCC GTGTGGGCCC
2251 CGACGCTGCC GACCCGCACG GCCGCGAGCA CTGGGCAGAG ATGCTGGCCA
2301 ATCCCCGCAA GCCCGTGGAG CACTGGCATC AGCTAGTGA GGAAGAACT
2351 GTGACCAGCT TCACAAAAGG CAGCAAAGGA CTATCAGAGA AAGAGAACTC
2401 CGAGTGAGGG GTCTGGCCTA GGCCCGGGAT CGGACCAGGC TCCCTCAGGA
20 2451 CCCCATCCTT TCCTGCCCCG ACCGTGAATT CATCTCCTTG AAGCCATAAC
2501 GTCCGAGCTG CTGGTGC GG GCGAGCCCTGG CCTAGGCTT CTAACCCCTG
2551 GAAGCGAGAG GATGAGAGGA GGCCGGCCCA GCTCCTTCTT TCAGGGTGGG
2601 GGTCAATTCAG CCTCCACTGT GTCTGTCTTT TCTTCCCTGG GGCTCCCCCT
2651 CGAGGCGAGG GGCCATGCAT GTCTGGGGGA CCCCTGCCCC CAAAAACCTT
25 2701 CTGTCTGTCT CTGTCTCTTT GCTGTTTGTG CAAGACTCAG TGTCCCGACC
2751 CTTGTTCTCG CCGTGAATGT CAATGGGCCA ATCCTCTCTG TCCTTTTCAGA
2801 CACACACACA CCTGTGTCCA CCCCTTCTGT TCGCCACACC CTGCGTCTGG
2851 CCGGTCCCCC CACTGCTGCT GCTATCAACG CCAGAATAAA CACACTCTGT
2901 GGGTCTCACT CCAAAAAAAA AAAAAAAAAA A
30

```

BLAST Results

```

35 Entry MMAB893_1 from database TREMBL:
   product: "synaptotagmin 3"; Mus musculus mRNA for synaptotagmin
   3,
   complete cds.
   Score = 1814, P = 5.7e-239, identities = 362/450, positives =
40 369/450,
   frame +2

```

45

Medline entries

```

96064733:
Fukuda M, Kojima T, Aruga J, Niinobe M, Mikoshiba K.; Functional
50 diversity of C2 domains of synaptotagmin family.
   Mutational analysis of inositol high polyphosphate binding
   domain. J
   Biol Chem 1995 Nov 3;270(44):26523-7

```

55

Peptide information for frame 2

ORF from 635 bp to 2404 bp; peptide length: 590

Category: strong similarity to known protein

5 Classification: Cell signaling/communication

Prosite motifs: C2_DOMAIN_1 (323-338)

C2_DOMAIN_1 (455-470)

```

10      1 MSGDYEDDLCL RRALILVSDL CARVRDADTN DRCQEFNDRI RGYPRGPDAD
      51 ISVSLLSVIV TFCGIVLLGV SLFVSWKLCW VPWRDKGGSA VGGGPLRKDL
     101 GPGVGLAGLV GGGGHHLAAG LGGHPLLGGP HHHAAHAHHP PFAELLEPGS
     151 LGGSDTPEPS YLDMDSYPEA AAAAVAAGVK PSQTSPELPS EGGAGSGLLL
     201 LPPSGGGLPS AQSHQQVTSI APTTRYPALP RPLTQQTLTS QPDPSSSEERP
15     251 PALPLPLPGG EEKAKLIGQI KPELYQGTGP GGRRSGGGPG SGEAGTGAPC
     301 GRISFALRYL YGSDQLVURI LQALDLPKQD SNGFSDPYVK IYLLPDRKKK
     351 FQTKVHRKTL NPVFNETFQF SVPLAELAQR KLHFSVYDFD RFSRHDLIGQ
     401 VVLNLLLELA EQPPDRPLWR DIVEGGSEKA DLGELNFSLC YLPTAGRLTV
     451 TTIKASNLKA MDLTGFSQPY VKASLISEGR RLKKRKTSIK KNTLNPTYNE
20     501 ALVFDVAPES VENVGLSIIV VDYDCIGHNE VIGVCRVGPD AADPHGREHW
     551 AEMLANPRKP VEHWHQLVEE KTVTSFTKGS KGLSEKENSE

```

25 BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_2013, frame 2

30 TREMBL:MMAB893_1 product: "synaptotagmin 3"; Mus musculus mRNA
for
synaptotagmin 3, complete cds., N = 2, Score = 1814, P = 1.1e-239

35 >TREMBL:MMAB893_1 product: "synaptotagmin 3"; Mus musculus mRNA
for
synaptotagmin 3, complete cds.
Length = 587

40 HSPs:

Score = 1814 (272.2 bits), Expect = 1.1e-239, Sum P(2) = 1.1e-239

45 Identities = 362/449 (80%), Positives = 369/449 (82%)

Query: 142 FAELLEPGSLGGSDTPEPSYLDMDSYPEXXXXXX-
XXGVKPSQXXXXXXXXXXXXXXXXXXXX 200

FAELLEPG LGGS+ PEPSYLDMDSYPE GVKPSQT

50 Sbjct: 143
FAELLEPGGLGGSELPEPSYLDMDSYPEAAVASVVAAGVKPSQTSPELPSEGGTGSGLLL 202

Query: 201
XXXXXXXXXXXXQSHQQVTSIAPTTTRYPALPRPLTQQTLTSQPDXXXXXXXXXXXXXXXXXXXX 260
QSHQQVTSIAPTTTRYPALPRPLTQQTLT+Q D

55 Sbjct: 203
LPPSGGGLPSAQSHQQVTSIAPTTTRYPALPRPLTQQTLTTQADPSTEERPPALPLPLPGG 262

Query: 261
XXKAKLIGQIKPELYQXXXXXXXXXXXXXXXXXXXXPCGRISFALRYLYGSDQLVVRI 320
KAKLIGQIKPELYQ
PCGRISFALRYLYGSDQLVVRI
5 Sbjct: 263 EEKAKLIGQIKPELYQGTGPGGRRGGGSGEAGA-----
PCGRISFALRYLYGSDQLVVRI 317

Query: 321
10 LQALDLPKDSNGFSDPYVKIYLLPDRKKKFQTKVHRKTLNPVFNETFQFSVPLAELAQR 380
LQALDLPKDSNGFSDPYVKIYLLPDRKKKFQTKVHRKTLNP+FNETFQFSVPLAELAQR
Sbjct: 318
LQALDLPKDSNGFSDPYVKIYLLPDRKKKFQTKVHRKTLNPIFNETFQFSVPLAELAQR 377

15 Query: 381
KLHFSVYDFDRFSRHDIGQVVLNLLLELAEQPPDRPLWRDIVEGGSEKADLGELNFSLC 440
KLHFSVYDFDRFSRHDIGQVVLNLLLELAEQPPDRPLWRDI+EGGSEKADLGELNFSLC
20 Sbjct: 378
KLHFSVYDFDRFSRHDIGQVVLNLLLELAEQPPDRPLWRDILEGGSEKADLGELNFSLC 437

Query: 441
YLPTAGRLTVTIKASNLKAMDLTGFSDPYVKASLISEGRRLKKRKTSIKKNTLNPTYNE 500
25 YLPTAGRLTVTIKASNLKAMDLTGFSDPYVKASLISEGRRLKKRKTSIKKNTLNPTYNE
Sbjct: 438
YLPTAGRLTVTIKASNLKAMDLTGFSDPYVKASLISEGRRLKKRKTSIKKNTLNPTYNE 497

Query: 501
30 ALVFDVAPESVENVGLSIAVVDYDCIGHNEVIGVCRVGPDAADPHGREHWAEMLANPRKP 560
ALVFDVAPESVENVGLSIAVVDYDCIGHNEVIGVCRVGP+AADPHGREHWAEMLANPRKP
Sbjct: 498
35 ALVFDVAPESVENVGLSIAVVDYDCIGHNEVIGVCRVGPEAADPHGREHWAEMLANPRKP 557

Query: 561 VEHWHQLVEEKT VTSFTKGSKGLSEKENSE 590
VEHWHQLVEEKT++SFTKG KGLSEKENSE
Sbjct: 558 VEHWHQLVEEKT LSSFTKG GKGGLSEKENSE 587

40 Score = 520 (78.0 bits), Expect = 1.1e-239, Sum P(2) = 1.1e-239
Identities = 98/100 (98%), Positives = 99/100 (99%)

Query: 1 MSGDYEDDLCCRALILVSDLCARVRDADTNDRCQEFND-
45 RIRGYPRGPDADISVSLLSVI 59
MSGDYEDDLCCRALILVSDLCARVRDADTNDRCQEFN+
RIRGYPRGPDADISVSLLSVI
Sbjct: 1
MSGDYEDDLCCRALILVSDLCARVRDADTNDRCQEFNELRIRGYPRGPDADISVSLLSVI 60

50 Query: 60 VTFCGIVLLGVSLFVSWKLCWVPWRDKGGSAVGGGGLRKD 99
VTFCGIVLLGVSLFVSWKLCWVPWRDKGGSAVGGGGLRKD
Sbjct: 61 VTFCGIVLLGVSLFVSWKLCWVPWRDKGGSAVGGGGLRKD 100

55 Pedant information for DKFZphamy2_2013, frame 2

Report for DKFZphamy2_2013.2

```

[LENGTH] 590
[MW] 63304.02
5 [pI] 6.16
[HOMOL] TREMBL:MMAB893_1 product: "synaptotagmin 3"; Mus
musculus mRNA for synaptotagmin 3, complete cds. 0.0
[FUNCAT] 99 unclassified proteins [S. cerevisiae, YML072c]
6e-10
10 [FUNCAT] 01.06.01 lipid, fatty-acid and sterol biosynthesis
[S. cerevisiae, YGR170w] 7e-06
[FUNCAT] 30.08 organization of golgi [S. cerevisiae, YGR170w]
7e-06
[BLOCKS] BL01224A N-acetyl-gamma-glutamyl-phosphate reductase
15 proteins
[BLOCKS] BL01013B Oxysterol-binding protein family proteins
[BLOCKS] PF01368B
[SCOP] d1a25a_ 2.6.1.2.2 C2 domain from protein kinase c
(beta) [Ra 2e-27
20 [SCOP] d1rsy_ 2.6.1.2.1 Synaptogamin I, first C2 domain
[Rat (Rattu 4e-43
[SCOP] d1rlw_ 2.6.1.1.2 A domain from cytosolic
phospholipase A2 [Huma 5e-12
[SCOP] d1qasb2 2.6.1.1.1 Phosphoinositide-specific
25 phospholipase C 4e-27
[PIRKW] phosphotransferase 7e-15
[PIRKW] duplication 6e-76
[PIRKW] synaptic vesicle 1e-167
[PIRKW] phorbol ester binding 2e-14
30 [PIRKW] zinc 2e-14
[PIRKW] transmembrane protein 0.0
[PIRKW] serine/threonine-specific protein kinase 7e-15
[PIRKW] membrane trafficking 0.0
[PIRKW] phospholipid binding 6e-76
35 [PIRKW] autophosphorylation 7e-15
[PIRKW] ATP 7e-15
[PIRKW] phosphoprotein 7e-15
[PIRKW] glycoprotein 1e-167
[PIRKW] calcium binding 5e-34
40 [PIRKW] alternative splicing 1e-10
[PIRKW] dimer 1e-75
[PIRKW] membrane protein 1e-167
[PIRKW] calmodulin binding 2e-74
[ESUPFAM] ras-specific GAP catalytic domain homology 1e-08
45 [ESUPFAM] protein kinase C zinc-binding repeat homology 7e-15
[ESUPFAM] protein kinase homology 7e-15
[ESUPFAM] protein kinase C alpha 7e-15
[ESUPFAM] HsC2 phosphatidylinositol 3-kinase 1e-09
[ESUPFAM] synaptotagmin 0.0
50 [ESUPFAM] PX domain homology 1e-09
[ESUPFAM] pleckstrin repeat homology 1e-08
[ESUPFAM] protein kinase C C2 region homology 0.0
[PROSITE] C2_DOMAIN_1 2
[PFAM] C2 domain
55 [KW] Irregular
[KW] 3D
[KW] LOW_COMPLEXITY 20.00 %

```

SEQ MSGDYEDDL CRRALILVSDLCARVRDADTNDRCQEFNDRIRGYPRGPDADISVSLLSVIV
 SEG
 lrsy-
 5

 SEQ TFCGIVLLGVSLFVSWKLCWVPWRDKGGSAVGGGGLRKDLGPGVGLAGLVGGGGHHLAAG
 SEGxxxxxxxxxxxxxxxxxxxxxxxxxxxx
 lrsy-
 10

 SEQ LGGHPLLGGPHHHAAHHPFAELLEPGSLGGSDTPEPSYLDMDSYPEAAAAVAAGVK
 SEG xxx
 lrsy-
 15

 SEQ PSQTSPELPSEGGAGSGLLLLPPSGGGLPSAQSHQVTS LAPTTRYPALPRPLTQQLTS
 SEGxx
 lrsy-
 20

 SEQ QPDPSSSEERPPALPLPLPGGEEKAKLIGQIKPELYQGTGPGGRRSGGGPGSGEAGTGAPC
 SEGxx
 lrsy-
 25

 SEQ GRISFALRYLYGSDQLVVRILQALDLPKDSNGFSDPYVKIYLLPDRKKKFQTKVHRKTL
 SEG
 lrsy-
 30 CEEEEEEEEETTTTEEEEEEEEEECCECCBTBTBBCEEEEEEEEEETTTTTEECCTTTBT

 SEQ NPVFNETFQFSVPLAELAQRKLHFSVYDFDRFSRHD LIGQVVLDNLLELAEQPPDRPLWR
 SEG
 lrsy-
 35 TEEEEEEEEECCHHHHCCEEEEEEEECTTTTCCEEEEE.....

 SEQ DIVEGGSEKADLGELNFSLCYLPTAGRLVTI IKASNLKAMDLTGFSDPYVKASLISEGR
 SEG
 lrsy-
 40

 SEQ RLKKRKTSIKKNTLNPTYNEALVFDVAPESVENVGLSIAVVDYDCIGHNEVIGVCRVGPD
 SEG
 lrsy-
 45

 SEQ AADPHGREHWAEMLANPRKPVEHWHQLVEEKT VTSFTKGSKGLSEKENSE
 SEG
 lrsy-
 50

Prosites for DKFZphamy2_2013-2

55	PS00499	323->339	C2_DOMAIN_1	PD0C00380
	PS00499	455->471	C2_DOMAIN_1	PD0C00380

Pfam for DKFZphamy2_2013.2

5 HMM_NAME C2 domain

HMM

*LtVrIIeARNLWkMDMnGfSDPYVKVdMdPdPkDtkKWKtKiWNNGLN
L+VRI++A +L+++D+NGFSDPYVK++++PD+K

10 KK++TK++++ LN

Query 316 LVVRILQALDLPKDSNGFSDPYVKIYLLPDRK--
KKFQTKVHRKT-LN 361

HMM

PVWNEEEFvFedIPyPdIqrkMLRFaVWDWDRFSRBDFIGHCi*

15

PV+N E+F+F +P+ +L+ + L+F+V+D+DRFSR+D+IG+++

Query 362 PVFN-ETFQFS-VPLAELAQRKLHFSVYDFDRFSRHDLIQVV
402

HMM

20

*LtVrIIeARNLWkMDMnGfSDPYVKVdMdPdPkDtkKWKtKiWNNGLN
LTV+II+A NL++MD +GFSDPYVK +++ +

+++KK+KT++++N+ LN

Query 448 -
LTVTIKASNLKAMDLTGFSDPYVKASLISEGRRLKKRKTSIKKNT-LN 495

25

HMM

PVWNEEEFvFedIPyPdIqrkMLRFaVWDWDRFSRBDFIGHCi*

P++N E +VF+ ++ ++ +++ L +AV D+D++++++IG+C+

Query 496 PTYN-EALVFD-VAPESVENVGLSIAVVYDYDCIGHNEVIGVCR
536

30

DKFZphamy2_7j5

group: differentiation/development

5

DKFZphamy2_7j5 encodes a novel 693 amino acid protein with similarity to Tspyl1 testis-specific Y-encoded-like protein of *Mus musculus*.

- 10 TSPY genes are arranged in clusters on the Y chromosome of many mammalian species. TSPY is believed to function in early spermatogenesis and is a candidate for GBY, the putative gonadoblastoma-inducing gene on the Y. The TSPY family forms part of a superfamily, TTSN, with autosomal representatives, highly
- 15 conserved in mammals and beyond.

The new protein can find application in studying the expression profile of testis- and brain-specific genes and diagnosis/therapy of malfunctioning male fertility.

20

HRIHFB221b

similarity to Y-linked Gene of *Mus musculus*

25

Sequenced by BMFZ

Locus: unknown

30 Insert length: 2819 bp

Poly A stretch at pos. 2800, polyadenylation signal at pos. 2779

```

35      1 AGGAGAGCTG GTTGCGTGAG TCTCCTCAGC TCTGCTTACC GGTGCGACTA
      51 GCGGCAGCGA CGCGGCTAAA AGCGAAGGGG CGAGTGCGAG TCCCCTGAGC
      101 TGTACGAACG CGGTCGCCAT GGACCGCCCA GATGAGGGGC CTCCGGCCAA
      151 GACCCGCCGC CTGAGCAGCT CCGAGTCTCC ACAGCGCGAC CCGCCCCCGC
      201 CGCCGCCGCC GCCGCCGCTC CTCCGACTGC CGCTGCCCTCC ACCCCAGCAG
      251 CGCCCGAGGC TCCAGGAGGA AACGGAGGCG GCACAGGTGC TGGCCGATAT
40      301 GAGGGGGGTG GGA CTGGGCC CCGCGCTGCC CCGCCGCCCT CCCTATGTCA
      351 TTCTCGAGGA GGGGGGGGATC CGCGCATACT TCACGCTCGG TGCTGAGTGT
      401 CCCGGCTGGG ATTCTACCAT CGAGTCGGGG TATGGGGAGG CCCCCCGGCC
      451 CACGGAGAGC CTGGAAGCAC TCCCCACTCC TGAGGCCTCG GGGGGGAGCC
      501 TGGAAATCGA TTTTCAGGTT GTACAGTCGA GCAGTTTGGG TGGAGAGGGG
45      551 GCCCTAGAAA CCTGTAGCGC AGTGGGGTGG GCGCCCCAGA GGTTAGTTGA
      601 CCCGAAGAGC AAGGAAGAGG CGATCATCAT AGTGGAGGAT GAGGATGAGG
      651 ATGAGCGGGA GAGTATGAGG AGCAGCAGGA GCGGCGGCGG CCGGCGGAGG
      701 AGGAAGCAGA GGAAGGTGAA GAGGGAAAGC AGAGAGAGAA ATGCCGAGAG
      751 GATGGAGAGC ATCCTGCAGG CACTGGAGGA TATTCAGCTG GATCTGGAGG
50      801 CAGTGAACAT CAAGGCAGGC AAAGCCTTCC TGCCTCTCAA GCGCAAGTTC
      851 ATCCAGATGC GAAGACCCTT CCTGGAGCGC AGAGACCTCA TCATCCAGCA
      901 TATCCCAGGC TTCTGGGTCA AAGCATTCCT CAACCACCCC AGAATTTCAA
      951 TTTTGATCAA CCGACGTGAT GAAGACATTT TCCGCTACTT GACCAATCTG
      1001 CAGGTACAGG ATCTCAGACA TATCTCCATG GGCTACAAAA TGAAGCTGTA
55      1051 CTTCCAGACT AACCCTACT TCACAAACAT GGTGATTGTC AAGGAGTTCC
      1101 AGCGCAACCG CTCAGGCCGG CTGGTGTCTC ACTCAACCCC AATCCGCTGG
      1151 CACCGGGGCC AGGAACCCCA GGCCCGTCGT CACGGGAACC AGGATGCGAG
      1201 CCACAGCTTT TTCAGCTGGT TCTCAAACCA TAGCCTCCCA GAGGCTGACA

```



```

1251 GGATTGCTGA GATTATCAAG AATGATCTGT GGGTTAACCC TCTACGCTAC
1301 TACCTGAGAG AAAGGGGGCTC CAGGATAAAG AGAAAGAAGC AAGAAATGAA
1351 GAAACGTAAA ACCAGGGGGCA GATGTGAGGT GGTGATCATG GAAGACGCCC
1401 CTGACTATTA TGCAGTGGAA GACATTTTCA GCGAGATCTC AGACATTGAT
5 1451 GAGACAATTC ATGACATCAA GATCTCTGAC TTCATGGAGA CCACCGACTA
1501 CTTGAGAGCC ACTGACAATG AGATAACTGA CATCAATGAG AACATCTGCG
1551 ACAGCGAGAA TCCTGACCAC AATGAGGTCC CCAACAACGA GACCACTGAT
1601 AACAACGAGA GTGCTGATGA CCACGAAACC ACTGACAACA ATGAGAGTGC
1651 AGATGACAAC AACGAGAATC CTGAAGACAA TAACAAGAAC ACTGATGACA
10 1701 ACGAAGAGAA CCTAACAAC AACGAGAACA CTTACGGCAA CAACTTCTTC
1751 AAAGGTGGCT TCTGGGGCAG CCATGGCAAC AACCAGGACA GCAGCGACAG
1801 TGACAATGAA GCAGATGAGG CCAGTGATGA TGAAGATAAT GTGGCAACG
1851 AAGGTGACAA TGAGGGCAGT GATGATGATG GCAATGAAGG TGACAATGAA
1901 GGCAGCGATG ATGACGACAG AGACATTGAG TACTATGAGA AAGTTATTGA
15 1951 AGACTTTGAC AAGGATCAGG CTGACTACGA GGACGTGATA GAGATCATCT
2001 CAGACGAATC AGTGGGAAGAA GAGGGCATTG AGGAAGGCAT CCAGCAAGAT
2051 GAGGACATCT ATGAGGAAGG AACTATGAG GAGGAAGGAA GTGAAGATGT
2101 CTGGGAAGAA GGGGAAGATT CGGACGACTC TGACCTAGAG GATGTGCTTC
2151 AGGTCCCCAA CGGTGGGGCC AATCCGGGGA AGAGGGGGAA AACC GGATAA
20 2201 GGGTTTTCCC CTTTTGGGGA TCACCTCTCT GTATCCCCCA CCCACTATCC
2251 CATTTGCCCT CCTCCTCAGC TAGGGCCACG CGGCCCCACA TTGCACTTCT
2301 GGGGGGTGAC CGACTTCGTA CACGGGTTTA AAGTTTATTT TTATGGTTTA
2351 GTCATTGCAG AGTTCTTATT TTGGGGGGAG GGAAGGGGGG CTAGTCCCCT
2401 TCTTTTGGCC CTCGCCCCC GCAGGCTTCT GTGTGCTGCT AACTGTATTT
25 2451 ATTGTGATGC CTTGGTCAGG GCCCCTCTAC CCACTTCTCC CAGTCAGTTG
2501 TGGCCCCAGC CCTCTCCCT GTGCTGTGTG GAGTGGACAC CCTGACCCCC
2551 GAAGCGGGGA GGGCCGCTGT GGCCTTCGTC ACAGCCGCGC AGTGCCCATG
2601 GAGGCGCTGC TGCCACCTTC CTCCTCCAAG TTCTTTCTCC ATCCCTCTCC
2651 TCTTCCCGCC GCGCCGCTAG CCCGCTCGG TGTCTATGCA AGGCCGCTTC
30 2701 GCCATTGCGG TATTCTTTGC GGTATTCTTG TCCCCGTCCC CCAGAAGGCT
2751 CGCCTCTCCC CGTGGACCCT GTTAATCCCA ATAAAATTCT GAGCAAGTTT
2801 AAAAAAAAAA AAAAAAAAAA

```

35 BLAST Results

No BLAST result

40 Medline entries

98399864:

```

45 Vogel T, Dittrich O, Mehraein Y, Dechend F, Schnieders F,
Schmidtke
J.; Murine and human TSPYL genes: novel members of the
TSPY-SET-NAP1L1 family. Cytogenet Cell Genet 1998;81(3-4):265-70

```

50

Peptide information for frame 2

55

ORF from 119 bp to 2197 bp; peptide length: 693
Category: similarity to known protein
Classification: unclassified


```

1  MDRPDEGPPA KTRRLSSSES PQRDP PPPPP PPPLRLPLP PPQQRPRRLQE
5  51 ETEAAQVLAD MRGVGLGPAL PPPPPYVILE EGGIRAYFTL GAECPGWDST
10 101 IESGYGEAPP PTESLEALPT PEASGGSLEI DFQVVQSSSF GGEGALETCS
15 151 AVGWAPQRLV DPKSKEEAI IVEDEDEDER ESMRSSRRRR RRRRKQKRV
20 201 KRESRERNAE RMESILQALE DIQLDLEAVN IKAGKAFLRL KRKFIQMRRP
25 251 FLERRDLIIQ HIPGFVVKAF LNHPRISILI NRRDEDIFRY LTNLQVQDLR
30 301 HISMGYKMKL YFQTNPYFTN MVIVKEFQRN RSGRLVSHST PIRWHRGQEP
35 351 QARRHGNQDA SHSFFSWFSN HSLPEADRIA EIIKNDLWVN PLRYYLREGR
10 401 SRIKRKKQEM KKRKTRGRCE VVIMEDAPDY YAVEDIFSEI SDIDETIHDI
45 451 KISDFMETTD YFETTDNEIT DINENICDSE NPDHNEVPNN ETTDNNESAD
50 501 DHETTDNNES ADDNNENPED NNKNTDDNEE NPNNNENTYG NNFFKGGFWG
55 551 SHGNNQDSSD SDNEADEASD DEDNDGNEG DNEGSDDDGNE GDNEGSDDDD
60 601 RDIEYYEKVI EDFDKDQADY EDVIEIISDE SVEEEGIEEG IQQDEDIYEE
15 651 GNYEEEGSED VWEEGEDSD SDLEDVLQVP NGWANPGKRG KTG

```

BLASTP hits

20

No BLASTP hits available

Alert BLASTP hits for DKFZphamy2_7j5, frame 2

```

25 TREMBL:AB015345_1 gene: "HRIHFB221b"; Homo sapiens HRIHFB221b
    mRNA,
    partial cds., N = 4, Score = 1393, P = 2.1e-165

30 TREMBL:HSDJ486I3_2 gene: "dJ486I3.2"; product: "dJ486I3.2
    (KIAA0721
    (NAP (Nucleosome Assembly Protein) domain containing protein))";
    Human
    DNA sequence from clone 486I3 on chromosome 6q22.1-22.3. Contains
    the
35 part of a gene for a novel protein, the gene for KIAA0721 (NAP
    (Nucleosome Assembly Protein) domain containing protein), the TSPYL
    gene
    for TSPY-like (testis specific protein, Y-linked like), and an
    RPS5
40 (40S Ribosomal Protein S5) pseudogene. Contains ESTs, STSs, GSSs
    and
    two putative CpG islands., N = 1, Score = 570, P = 3.4e-55

45 >TREMBL:AB015345_1 gene: "HRIHFB221b"; Homo sapiens HRIHFB221b
    mRNA,
    partial cds.
        Length = 486

50 HSPs:

    Score = 1393 (209.0 bits), Expect = 2.1e-165, Sum P(4) = 2.1e-
    165
    Identities = 268/295 (90%), Positives = 268/295 (90%)

55 Query: 208
    NAERMESILQALEDIQLDLEAVNIKAGKAFLRLKRKFIQMRRPFLERRDLIIQHIPGFV 267

```

NAERMESILQALEDIQLDLEAVNIKAGKAFLRLKRKFIQMRRPFLERRDLIIQHIPGFUV

Sbjct: 1

NAERMESILQALEDIQLDLEAVNIKAGKAFLRLKRKFIQMRRPFLERRDLIIQHIPGFUV 60

5

Query: 268

KAFLNHPRISILINRRDEDIFRYLTNLQVQDLRHISMGYKMKLYFQTNPYFTNMVIVKEF 327

KAFLNHPRISILINRRDEDIFRYLTNLQVQDLRHISMGYKMKLYFQTNPYFTNMVIVKEF

10 Sbjct: 61

KAFLNHPRISILINRRDEDIFRYLTNLQVQDLRHISMGYKMKLYFQTNPYFTNMVIVKEF 120

Query: 328

QRNRSGRLVSHSTPIRWHRGQEPQARRHGNQDAXXXXXXXXXXXXXLPEADRIAEEIKNDL 387

15

QRNRSGRLVSHSTPIRWHRGQEPQARRHGNQDA

LPEADRIAEEIKNDL

Sbjct: 121

QRNRSGRLVSHSTPIRWHRGQEPQARRHGNQDASHSFFSWFSNHSLEADRIAEEIKNDL 180

20

Query: 388

WVNPLRYLLRERGSXXXXXXXXXXXXXXXXXGRCEVVIMEDAPDYYAVEDIFSEISDIDETI 447

WVNPLRYLLRERGS

GRCEVVIMEDAPDYYAVEDIFSEISDIDETI

Sbjct: 181

25

WVNPLRYLLRERGSRIKRKKQEMKKRKTRGRCEVVIMEDAPDYYAVEDIFSEISDIDETI 240

Query: 448

HDIKISDFMETTDYFETTDNEITDINENICDSENPDPHNEVPNNETTDNNESADDH 502

30

HDIKISDFMETTDYFETTDNEITDINENICDSENPDPHNEVPNNETTDNNESADDH

Sbjct: 241

HDIKISDFMETTDYFETTDNEITDINENICDSENPDPHNEVPNNETTDNNESADDH 295

Score = 117 (17.6 bits), Expect = 9.0e-19, Sum P(4) = 9.0e-19

35

Identities = 32/77 (41%), Positives = 44/77 (57%)

Query: 426

DAPDYYAVEDIFSEISDIDETIHDIKISDFMETTDYFETTDNEITDINENICDSENPDPHN 485

+DY+ D +EI+DI+E I D E D+ E +NE TD NE+

40

D E D+N

Sbjct: 250 ETTDYFETTD--NEITDINENICD-----

SENPDPHNEVPNNETTDNNESADDHETTDNN 301

Query: 486 EVP--NNETT--DNNESADDH 502

45

E NNE DNN++ DD+

Sbjct: 302 ESADDNNENPEDNNKNTDDN 321

Score = 94 (14.1 bits), Expect = 2.1e-165, Sum P(4) = 2.1e-165

50

Identities = 16/16 (100%), Positives = 16/16 (100%)

Query: 678 QVPNGWANPGKRGKTG 693

QVPNGWANPGKRGKTG

Sbjct: 471 QVPNGWANPGKRGKTG 486

55

Score = 90 (13.5 bits), Expect = 9.9e-16, Sum P(4) = 9.9e-16

Identities = 34/85 (40%), Positives = 45/85 (52%)

Query: 426 DAPDYYAVEDIFSEISDIDETIHDIKISDFME-----TTDYFETTDN-
 EITDINENICDS 479
 + DY+ D +EI+DI+E I D + D E TTD E+ D+ E TD

NE+ D+

5 Sbjct: 250 ETTDYFETTD--
 NEITDINENICDSENPDPHNEVPNNETTDDNNEASADDHETTDNNEASADDN 307

Query: 480 -ENPDHN-----EVPNN-ETTDNN 496
 ENP+ N E PNN E T N

10 Sbjct: 308 NENPEDNNKNTDDNEENPNNNENTYGN 334

Score = 87 (13.1 bits), Expect = 2.1e-165, Sum P(4) = 2.1e-165
 Identities = 14/14 (100%), Positives = 14/14 (100%)

15 Query: 543 FFKGGFWGSHGNNQ 556

FFKGGFWGSHGNNQ

Sbjct: 336 FFKGGFWGSHGNNQ 349

20 Score = 85 (12.8 bits), Expect = 2.1e-165, Sum P(4) = 2.1e-165
 Identities = 16/18 (88%), Positives = 17/18 (94%)

Query: 601 RDIEYYEKVIEDFDKDQA 618

RDIEYYEK IEDFD+DQA

25 Sbjct: 394 RDIEYYEKGIEDFDRDQA 411

Score = 60 (9.0 bits), Expect = 5.3e-03, Sum P(4) = 5.3e-03
 Identities = 21/66 (31%), Positives = 33/66 (50%)

30 Query: 426 DAPDYYAVEDIFSEISDIDETIHD-IKIS-
 DFMETTDYFETTDNEITDINENICDSENPDP 483

D DY V +I S+ S +E I + I+ D E +Y E ++ + E+

DS+ D

Sbjct: 409

DQADYEDVIEIISDESVEEEGIEEGIQQDEDIYEEGNYEEEGSEDVWEEGEDSDDDSDLED 468

35

Query: 484 HNEVPN 489

+VPN

Sbjct: 469 VLQVPN 474

40 Score = 49 (7.4 bits), Expect = 1.4e-06, Sum P(4) = 1.4e-06
 Identities = 12/35 (34%), Positives = 21/35 (60%)

Query: 463 ETTDNEITDINENICDSENPDPHNEVPNNETTDDNNE 497

E +D+E D NE + + D NE +NE +D+++

45 Sbjct: 360 EASDDEDNDGNEGDNEGSDDDGNE-GDNEGSDDDD 393

Score = 42 (6.3 bits), Expect = 7.2e-06, Sum P(4) = 7.2e-06
 Identities = 11/37 (29%), Positives = 18/37 (48%)

50 Query: 465 TDNEITDINENICDSENPDPHNEVPNNETTDDNNEASADD 501

+DNE + + D E+ D NE N + D+ D+

Sbjct: 354 SDNEADEAS---DDEDNDGNEGDNEGSDDDGNEGDN 386

55

Pedant information for DKFZphamy2_7j5, frame 2

Report for DKFZphamy2_7j5-2

-208-

SEQ SHSFFSWFSNHSLPEADRIAEIKNDLWVNPLRYYLRRERGSRIKRKKQEMKKRKTRGRCE
SEG xxx
PRD cccccccccccccccccchhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhccceeeeeecccccc

5 SEQ VVIMEDAPDYAVEDIFSEISDIDETIHDIKISDFMETTDYFETTDNEITDINENICDSE
SEG
PRD eeccccccccceehhhhhhhhhhhccccccccceeeccccccccccccchhhhhhhhhcccccc

10 SEQ NPDHNEVPNNETTNNESADDHETTDNNESADDNNENPEDNNKNTDDNEENPNNNENTYG
SEGxx
PRD cccccccccceeeccccccccccccccccccccchhhhhccccccccceeecccccccccccccccc

15 SEQ NNFFKGGFWGSHGNNQDSSDSNEADEASDDEDNDGNEGDNEGSDDDGNEGDNEGSDDDD
SEG xx.....xx
PRD ccc

20 SEQ RDIEYYEKVIEDFDKQADYEDVIEIISDESVEEEGIEEGIQQDEDIYEEGNYEEEGSED
SEGxx
PRD cchhhhhhhhhhhhhccccchhhhhheeeccccccccccccccccccccceeeccccccccccce

25 SEQ VWEEGEDSDSDLEDVLQVPNGWANPGKRGKTG
SEG xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
PRD eeccccccccccccceeecccccccccccccccccc

(No Prosite data available for DKFZphamy2_7j5.2)

(No Pfam data available for DKFZphamy2_7j5.2)

30 Pedant information for DKFZphamy2_7j5, frame 3

Report for DKFZphamy2_7j5.3

35 [LENGTH] 150
[MW] 16810.69
[pI] 12.88
40 [BLOCKS] PRODB08A
[KW] All_Alpha
[KW] LOW_COMPLEXITY 61.33 %

45 SEQ MRTSATARILTTMRSPTTRPLITTRVLMTTKPLTTMRVQMTTTRILKTITRTLMTTKRTL
SEGxxxxxxxxxxxxxxxxxxxxxxxxxxxx
PRD ccchhhhhhhhhhhccccccccccccceeeccccccccchhhhhhhhhhhhhhhhhhhhhcccccc

50 SEQ TTTRTLTATTSSKVASGAAMATTRTAATVTMKQMRPVMMKIMMATKVTMRAVMMAMKVT
SEG xxxxxxxxxxx-.....xx
PRD cccccceeeccccccccchhhhhhhhhhhhhhhhhhhchhhhhhhhhhhhhhhhhhhhhhhhhhhhh

55 SEQ MKAAMTTETLSTMRKLLKTLTRIRLTTRT
SEG xxxxxxxx-xxxxxxxxxxxxxxxxxxxxxxxxxxxx
PRD hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcc

(No Prosite data available for DKFZphamy2_7j5.3)

(No Pfam data available for DKFZphamy2_7j5.3)

DKFZphfbr2_78c12

5

group: nucleic acid management

10 DKFZphfbr2_78c12 encodes a novel 528 amino acid protein with high similarity to glutamyl-tRNA (Gln) amidotransferase subunit A of the hyperthermophilic bacterium Aquifex aeolicus.

15 The novel protein contains one ATP/GTP-binding site motif A (P-loop). This loop interacts with one of the phosphate groups of a A or G nucleotide. It is found in numerous ATP- or GTP-binding proteins, such as ATP synthase alpha and beta subunits, Myosin heavy chains, Kinesin heavy chains and kinesin-like proteins, Dynamins and dynamin-like proteins, several kinases, DNA and RNA
20 helicases, GTP-binding elongation factors and the Ras family of GTP-binding proteins. The protein seems to be expressed ubiquitously.

25 The new protein can find application in the modulation of translational pathways.

30 similarity to glutamyl-tRNA (Gln) amidotransferase subunit A (Aquifex aeolicus)

Sequenced by MediGenomix

35 Locus: /map="b8b.3 cR from top of Chrb linkage group"

Insert length: 3244 bp

Poly A stretch at pos. 3222, polyadenylation signal at pos. 3204

40 1 AGTGACAATT AAAGATGGCT GCGCCCATGT AACATCACTA GCGACCGGTG
51 ACCTCTTTTT CCCCCTTGCC TGGCTCCTGT GGTGGCAGGC TGGGCACGAG
101 GACCATGCTG GGCCGGAGCC TCCGAGAAGT TTCTGCGGCA CTGAAACAAG
151 GCCAAATTAC ACCAACAGAG CTCTGTCAAA AATGTCTCTC TCTTATCAAG
201 AAGGCCAAGT TTCTAAATGC CTACATTACT GTGTCAGAAG AGGTGGCCTT
45 251 AAAACAAGCT GAAGAATCAG AAAAGAGATA TAAGAATGGA CAGTCACTTG
301 GGGATTTAGA TGGAAATCCT ATTGCAGTAA AAGACAATTT CAGCACTTCT
351 GGCATTGAGA CAACATGTGC ATCAAATATG CTGAAAGGTT ATATACCACC
401 TTATAATGCT ACAGTAGTTC AGAAGTTGTT GGATCAGGGA GCTCTACTAA
451 TGGGAAAAAC AAATTTAGAT GAGTTTGCTA TGGGATCTGG GAGCACAGAT
50 501 GGTGTATTTG GACCACTTAA AAACCCCTGG AGTTATTCAA AACGATATAG
551 AGAAAAGAGG AAGCAGAATC CCCACAGCGA GAATGAAGAT TCAGACTGGC
601 TGATAACTGG AGGAAGCCCA GGTGGGAGTG CAGCTGCTGT ATCGGCGTTC
651 ACATGCTACG CGGCTTTAGG ATCAGATACA GGAGGATCGA CCAGAAATCC
701 TGCTGCCCCA TGTGGGCTTG TTGGTTTCAA ACCAAGCTAT GGCTTAGTTT
55 751 CCCGTCATGG TCTCATTCCT CTGGTGAATT CGATGGATGT GCCAGGAATC
801 TTAACCAGAT GTGTGGATGA TGCAGCAATT GTGTTGGGTG CACTGGCCGG
851 ACCTGACCCC AGGGACTCTA CCACAGTACA TGAACCTATT AATAAACCAT
901 TCATGCTTCC CAGTTTGGCA GATGTGAGCA AACTATGTAT AGGAATTCCA

5 951 AAGGAATATC TTGTACCGGA ATTATCAAGT GAAGTACAGT CTCTTTGGTC
1001 CAAAGCTGCT GACCTCTTTG AGTCTGAGGG GGCCAAAGTA ATTGAAGTAT
1051 CCCTTCTCTCA CACCAGTTAT TCAATTGTCT GCTACCATGT ATTGTGCACA
1101 TCAGAAGTGG CATCGAATAT GGCAAGATTT GATGGGCTAC AATATGGTCA
1151 CAGATGTGAC ATTGATGTGT CCACTGAAGC CATGTATGCT GCAACCAGAC
1201 GAGAAGGATT TAATGATGTG GTGAGAGGAA GAATTCTCTC AGGAAACTTT
1251 TTCTTATTAA AAGAAAAC TAAGAAATTAT TTTGTCAAAG CACAGAAAGT
1301 GAGACGCCTC ATTGCTAATG ACTTTGTAAA TGCTTTTAAAC TCTGGAGTAG
1351 ATGTCTTGCT AACTCCCAAC ACCTTGAGTG AGGCAGTACC ATACTTGGAG
10 1401 TTCATCAAAG AGGACAACAG AACCCGAAGT GCCCAGGATG ATATTTTTAC
1451 ACAAGCTGTA AATATGGCAG GATTGCCAGC AGTGAGTATC CCTGTTGCAC
1501 TCTCAAACCA AGGGTTGCCA ATAGGACTGC AGTTTATTGG ACGTGCGTTT
1551 TGTGACCAGC AGCTTCTTAC AGTAGCCAAA TGGTTTGAAA AACAAGTACA
1601 GTTTCCTGTT ATTCAACTTC AAGAACTCAT GGATGATTGT TCAGCAGTCC
15 1651 TTGAAAATGA AAAGTTAGCC TCTGTCTCTC TAAAACAGTA AACATATCTT
1701 ACAAATTAATA ATGACTTTTA GGCTGGGTGC AGTGGCTCAC ACCTGTAATC
1751 CCAGCACTTT GGGAGGCCAA GGCAGCGGA TCATGAGGTC AGAAGATCTA
1801 GAACAGCCTG GTCAACATGG TGAAACCCCG TCTCTACTAA AAATACAAAA
1851 ATTAGCCAGG CTTAGTGGCG GGCATCTGTA GTCCCAGCTA CTCAGGAGGC
20 1901 TGAGGCAGGA GAATCACTTG AACCTGGAG GTGGAGGTTG CAGTGAGCCG
1951 AGATCACTGC ACTGCACTGC ACTCCAGCCT GGGTGACAAA GCAAGACTGT
2001 GTCTCAAAAT AAATAAATAA AATAAAATAA AATGACGTAC AGAGACTCTA
2051 TATTCTAGAG AGTCAAATGG TCTTGCTCAA TTCTTGTAAT TAGGTTCTTG
2101 TTAATACAGT CATTCCATGG AATTACTTTT TAAAATTCTT GTGACAATTA
25 2151 ATAATAAATA ACGTGTGAGC ATTTAGTAAG CATCCACTAA GTGTACAATA
2201 CTTCTACAAT AACACAAGAT ACCTGTTCTT CAAAGACAAT GCATTCTGCC
2251 ATAATGTTCA TTAAAGAGTT TACAGTAAAA ATAAGATTAG GGATAAACTT
2301 CTCAAAAATT GTACATCTGT GTAACATAAG CACTAACAAA AACATGAATA
2351 GTCCTTCTAG AGGTAACCTG GATAGCCTAG GCAGGCAACT TATCATGTGG
30 2401 TGAAGGCCGC CTCAGGGGTT GTTAAAAATG CACAGAAACA ATTGAGTGGC
2451 ATTATTGGCT TCTGAGCGCT GAGCAGAGCA GGTGGAAGAG GAACTTTGAG
2501 CACAGGAGGA AATGCAACCA GTCAGGGCCC AGAATCATGC AAATCTCAGG
2551 GGTATGCCTC TCTGGGGAGG AGCTCCACTT GCAGGGACTC CTTTTATTTT
2601 CCTAAGAAAG AGCTGAAATG ACTGAGAACT TTCCTTTCCT CCTTAGAGTT
35 2651 ACAATTTTAC TTCTGCTATT CCGGAGCCCA TGCCTAGAAG CCAGAACAAAC
2701 TCCATGTTAC ACTGAGTTCA TGCTCCTATT TACTGATCAC AAATGAGCTC
2751 ATTAATGTCA TCGAAACATT TATTGTAACC TAACAGACCA TCACAGATTG
2801 GAAACTTGGT AGATAGCAGA GCATGGTATT AGTGAAAAAG GTTCAAAATA
2851 CACAAGTAAC ATACACTCTG AAAAAACATG AGATAATTTG CTGATGAAGC
40 2901 AGAAGAGGGG ATGCGCATGG CAAGAACTTG CCTTACCCCA GATTCTCTAT
2951 ATCTCATGGT TTCCTTTTCC TCTTGACTGT CTTTACGAGT GTTTTTTATT
3001 TGGGACCCCTC GAGCCCAGAG ATATTAATGG ATATCTGTAT TCAATATTTG
3051 ACAAATCTA ATGGAAACCA TCCATTTACT CATGATAAGG CTTCACTACT
3101 GGATTTCTGT GTCTTCACTA GAACACCATT GTCATCTCAT ATTGATCAGG
45 3151 TATTTTAATC TAGCACTTAC ATATTGTTGA TAAATGAAAG CTGAATTGTT
3201 ACTTAATAAA TTCACTTTGT TTAGCAAAAA AAAAAAAAAA AAAA

BLAST Results

50

No BLAST result

55

Medline entries

No Medline entry

Peptide information for frame 3

5

ORF from 105 bp to 1688 bp; peptide length: 528

Category: similarity to known protein

Classification: Protein management

10 Prosite motifs: ATP_GTP_A (112-119)

1 MLGRSLREVS AALKQGGQITP TELCQKCLSL IKKAKFLNAY ITVSEEVALK
 51 QAESEKRYK NGQSLGDLDG IPIAVKDNFS TSGIETTCAS NMLKGYIPPY
 15 101 NATVVQKLLD QGALLMGKTN LDEFAMGSGS TDGVFGPVKN PWSYSKRYRE
 151 KRKQNPHESEN EDSQWLITGG SPGGSAAAVS AFTCYAALGS DTGGSTRNPA
 201 AHCGLVGFKP SYGLVSRHGL IPLVNSMDVP GILTRCVDDA AIVLGALAGP
 251 DPRDSTTVHE PINKPFMLPS LADVSKLCIG IPKEYLVPPEL SSEVQSLWSK
 301 AADLFESEGA KVIEVSLPHT SYSIVCYHVL CTSEVASNMA RFDGLQYGHR
 20 351 CDIDVSTEAM YAATRREGFN DVVRGRILSG NFFLLKENYE NYFVKAQKVR
 401 RLIANDFVNA FNSGVDVLLT PTTLSEAVPY LEFIKEDNRT RSAQDDIFTQ
 451 AVNMAGLPAV SIPVALSNQG LPIGLQFIGR AFCDAQLLTV AKWFEKQVQF
 501 PVIQLQELMD DCSAVLENEK LASVSLKQ

25

BLASTP hits

No BLASTP hits available

30

Alert BLASTP hits for DKFZphfbr2_78cl2, frame 3

PIR:F70322 glutamyl-tRNA (Gln) amidotransferase subunit A -
 Aquifex

35 aeolicus, N = 2, Score = 620, P = 4.3e-89

>PIR:F70322 glutamyl-tRNA (Gln) amidotransferase subunit A -
 Aquifex

40 aeolicus

Length = 478

HSPs:

45 Score = 620 (93.0 bits), Expect = 4.3e-89, Sum P(2) = 4.3e-89
 Identities = 135/319 (42%), Positives = 195/319 (61%)

Query: 187

ALGSDTGGSTRNPAAHCGLVGFKPSYGLVSRHGLIPLVNSMDVP GILTRCVDDAAIVLGA 246
 +LGSDTGGG R PA+ CG++G KP+YG VSR+GL+ +S+D G+ R

50

+D A+VL

Sbjct: 163

SLGSDTGGGIRQPASFCGVIGIKPTYGRVSRVGLVAFASSLDQIGVFGRRTEDVALVLEV 222

55

Query: 247

LAGPDPRDSTTVHEPINKPFMLPSLADVSKLCIGIPKEYLVPPELSSEVQSLWSKAADLFE 306
 ++G D +DST+ P+ + + +V L IG+PKE+ EL +V+ +

E

Sbjct: 223 ISGWDEKDDSTSAKVPVPE-
WSEEVKKEVKGLKIGLPKEFFEYELQPPVKEAFENFIKELE 281

Query: 307

5 SEGAKVIEVSLPHTSYSIVCYHVLCTSEVASNMARFDGLQYGHRCDDIDVSTEAMYAATTR 366
EG ++ EVSLPH YSI Y+++ SE +SN+AR+DG++YG+R

MYA TR

Sbjct: 282

10 KEGFEIKEVSLPHVKYSIPTYYIIAPSEASSNLARYDGVRYGYRAKEYKDIFEMYARTRD 341

Query: 367

EGFNDVVRGRILSGNFFLLKENYENYFVKAQKVRRLIANDFVNAFNSGVDVLLTPTTLSE 426
EGF V+ RI+ G F L Y+ Y++KAQKVRRLI NDF+ AF VDV+
+PTT

15 Sbjct: 342 EGFGEVKKRRIMLGTALSAGYYDAYYLKAQKVRRLITNDFLKAFFEE-
VDVIASPTT--P 398

Query: 427

20 AVPYLEFIKEDNRTSAQDDIFTQAVNMAGLPAVSIPVALSNQGLPIGLQFIGRAFCDDQ 486
+P+ + +N DI T N+AGLPA+SIP+A + GLP+G Q

IG+ + +

Sbjct: 399 TLPFKFGERLENPIEMYLSDILTPANLAGLPAISIPIAWKD-
GLPVGGLIGKHWDETT 457

25 Query: 487 LLTVAK-WFEKQVQFPVIQL 505

LL ++ W +K + I L

Sbjct: 458 LLQISYLWEQKFKHYEKIPL 477

30 Score = 289 (43.4 bits), Expect = 4.3e-89, Sum P(2) = 4.3e-89
Identities = 64/143 (44%), Positives = 90/143 (62%)

Query: 4 RSLREVSAALKQGGITPTELCQKCLSLIKKAKF-
LNAYITVSEEVALKQAESEKRYKNG 62

35 ++R +SL E+ LK+G+++P E+ + + + AYIT ALKQAE

Sbjct: 5

KSLSELRELLKRGEVSPKEVVESFYDRYNQTEEVKAYITPLYGKALKQAESLKER---- 60

Query: 63

40 QSLGDLGDIPIAVKDNFSTSGIETTCASNMLKGYIPPYNATVVQKLLDQGALLMGKTNLD 122
L L GIPIAVKDN G +TTCAS +L+ ++ PY+ATV+++L

GAL++GKTNLD

Sbjct: 61 -EL-

45 PLFGIPIAVKDNILVEGEKTTCAKILENFVAPYDATVIERLKKAGALIVGKTNLD 118

Query: 123 EFAMGSGSTDGVFGPVKNPWSYSK 146

EFAMGS + F P KNPW +

Sbjct: 119 EFAMGSSTEYSAFFPTKNPWDLER 142

50

Pedant information for DKFZphfbr2_78c12, frame 3

Report for DKFZphfbr2_78c12.3

55

[[LENGTH]] 528
[[MW]] 57468.78

-214-

```
SEQ  AFCDQQLLTVAKWFEKQVQFPVIQLQELMDDCSAVLENEKLASVSLKQ
SEG  .....
PRD  cccchhhhhhhhhhhhhhhhhhhheeehhhhhhheeecccccceeeccc
```

PS00017 112->120 ATP_GTP_A PDC00017

DKFZphfbr2_78d18

DKFZphfbr2_78d1a encodes a novel 535 amino acid protein with weak similarity to a human putative mitogen-activated protein kinase kinase kinase.

The new protein can find application in studying the expression profile of brain-specific genes.

(Homo sapiens)

Locus: unknown

Poly A stretch at pos. 2138, polyadenylation signal at pos. 2117

1	ATCCGGGGCC	CCGGAACCCG	AGCTGGAGCT	GAAGCGCAGG	CTGCGGGGCG
51	CGGAGTCGGG	AGTGCAGGCC	TGAGTGTTC	TTCCAGCATG	TCGGAGGGGG
101	AGTCCAGAC	AGTACTTAGC	AGTGGCTCAG	ACCCAAAGGT	AGAATCCTCA
151	TCTTCAGCCC	CTGGCCTGAC	ATCAGTGTCA	CCTCCTGTGA	CCTCCACAAC
201	CTCAGCTGCT	TCCCCAGAGG	AAGAAGAAGA	AAGTGAAGAT	GAGTCTGAGA
251	TTTTGGAAGA	GTGCCCCTGT	GGGCGCTGGC	AGAAGAGGCG	AGAAGAGGTT
301	AATCAACGGA	ATGTACCAGG	TATTGCACAGT	GCATACCTGG	CCATGGATAC
351	AGAGCAAGGT	GTAGAGGTTG	TGTGGAATGA	GGTACAGTTC	TCTGAACGCA
401	AGAACTACAA	GCTGCAGGAG	GAAAAGGTTT	GTGCTGTGTT	TGATAATCTG

451 ATTCAATTGG AGCATCTTAA CATTGTAAAG TTTCACAAAT ATTGGGCTGA
501 CATTAAAGAG AACAAAGGCCA GGGTCATTTT TATCACAGAA TACATGTCAT
551 CTGGGAGTCT GAAGCAATTT CTGAAGAAGA CCAAAAAGAA CCACAAGACG
601 ATGAATGAAA AGGCATGGAA GCGTTGGTGC ACACAAATCC TCTCTGCCCT
5 651 AAGCTACCTG CACTCCTGTG ACCCCCCCAT CATCCATGGG AACCTGACCT
701 GTGACACCAT CTTTCATCCAG CACAACGGAC TCATCAAGAT TGGCTCTGTG
751 GCTCCTGACA CTATCAACAA TCATGTGAAG ACTTGTGAG AAGAGCAGAA
801 GAATCTACAC TTCTTTGCAC CAGAGTATGG AGAAGTCACT AATGTGACAA
851 CAGCAGTGGG CATCTACTCC TTTGGCATGT GTGCACTGGA GATGGCAGTG
10 901 CTGGAGATTG AGGGCAATGG AGAGTCCTCA TATGTGCCAC AGGAAGCCAT
951 CAGCAGTGCC ATCCAGCTTC TAGAAGACCC ATTACAGAGG GAGTTCATTG
1001 AAAAGTGCCT GCAGTCTGAG CCTGCTCGCA GACCAACAGC CAGAGAACTC
1051 CTGTTCCACC CAGCATTGTT TGAAGTGCCC TCGCTCAAAAC TCCTTGCGGC
1101 CCACTGCATT GTGGGACACC AACACATGAT CCCAGAGAAC GCTCTAGAGG
15 1151 AGATCACCAA AAACATGGAT ACTAGTGCCG TACTGGCTGA AATCCCTGCA
1201 GGACCAGGAA GAGAACCAGT TCAGACTTTG TACTCTCAGT CACCAGCTCT
1251 GGAATTAGAT AAATTCCTTG AAGATGTGAG GAATGGGATC TATCCTCTGA
1301 CAGCCTTTGG GCTGCCTCGG CCCCAGCAGC CACAGCAGGA GGAGGTGACA
1351 TCACCTGTG TGGCCCCCTC TGTCAAGACT CCGACACCTG AACCAGCTGA
20 1401 GGTGGAGACT CGCAAGGTGG TGCTGATGCA GTGCAACATT GAGTCGGTGG
1451 AGGAGGGAGT CAAACACCAC CTGACACTTC TGCTGAAGTT GGAGGACAAA
1501 CTGAACCGGC ACCTGAGCTG TGACCTGATG CCAAATGAGA ATATCCCCGA
1551 GTTGGCGGCT GAGCTGGTGC AGCTGGGCTT CATTAGTGAG GCTGACCAGA
1601 GCCGGTTGAC TTCTCTGCTA GAAGAGACCT TGAACAAGTT CAATTTTGCC
25 1651 AGGAACAGTA CCCTCAACTC AGCCGCTGTC ACCGTCTCCT CTTAGAGCTC
1701 ACTCGGGCCA GGGCCTGATC TGCGCTGTGG CTGTCCCTGG ACGTGCTGCA
1751 GCCCTCCTGT CCCTTCCCCC CAGTCAGTAT TACCCTGTGA AGCCCCCTCC
1801 CTCCTTTATT ATTCAGGAGG GCTGGGGGGG CTCCCTGGTT CTGAGCATCA
1851 TCCTTTCCCC TCCCCTCTCT TCCTCCCCTC TGCACTTTGT TTACTTGTTT
30 1901 TGCACAGACG TGGGCCTGGG CCTTCTCAGC AGCCGCCTTC TAGTTGGGGG
1951 CTAGTCGCTG ATCTGCCGGC TCCCGCCAG CCTGTGTGGA AAGGAGGCCC
2001 ACGGGCACTA GGGGAGCCGA ATTCTACAAT CCCGCTGGGG CGGCCGGGGC
2051 GGGAGAGAAA GGTGGTGCTG CAGTGGTGGC CCTGGGGGGC CATTGATTTC
2101 GCCTCAGTTG CTGCTGTAAT AAAAGTCTAC TTTTGGCCAA AAAAAAAAAA
35 2151 AAAAAAAAAA

BLAST Results

40

No BLAST result

Medline entries

45

No Medline entry

50

Peptide information for frame 1

55

ORF from 88 bp to 1692 bp; peptide length: 535

Category: similarity to unknown protein

Classification: Protein management

1 MSEGESQTVL SSGSDPKVES SSSAPGLTSV SPPVTSTTSA ASPEEEEESE

```

51 DESEILEESP CGRWQKRREE VNQRNVP GID SAYLAM DTEE GVEVVWNEVQ
101 FSERK NYKLQ EEKVR AVFDN LIQLEHLNIV KFHKYWADIK ENKARVIFIT
151 EYMSSGSLKQ FLKKT KKNHK TMNEKA WKRW CTQILSALSY LHSCDPPIIH
201 GNLTCDTIFI QHNGLIKIGS VAPDTINNHV KTCREEQKNL HFFAPEYGEV
5 251 TNVTTAVDIY SFGMCALEMA VLEIQNGES SYVPQEAISS AIQLLEDPLQ
301 REFIQKCLQS EPARRPTARE LLFHPALFEV PSLKLLAAHC IVGHQHMIPE
351 NALEEITKNM DTSAVLAEIP AGPGREP VQT LYSQSPAEL DKFLEDVRNG
401 IYPLTAFGLP RPQQPQQEEV TSPVVPPSVK TPTPEPAEVE TRKVVLMOCN
451 IESVEEGVKH HLTLLKLED KLNRLHSCDL MPNENIPELA AELVQLGFIS
10 501 EADQSR L TSL LEETLNKFNF ARNSTLNSAA VTVSS

```

BLASTP hits

15

No BLASTP hits available

Alert BLASTP hits for DKFZphfbr2_78d18, frame 1

```

20 TREMBL:AC009465_14 gene: "T9J14.14"; product: "putative mitogen
activated protein kinase kinase"; Arabidopsis thaliana
chromosome III
BAC T9J14 genomic sequence, complete sequence., N = 1, Score =
25 372, P =
1.9e-33

TREMBL:AF145690_1 gene: "BcDNA.LD28657"; product:
"BcDNA.LD28657";
Drosophila melanogaster clone LD28657 BcDNA.LD28657
30 (BcDNA.LD28657)
mRNA, complete cds., N = 1, Score = 1140, P = 1.3e-115

PIR:T02951 probable mitogen activated protein kinase - rice, N =
35 1,
Score = 391, P = 1.4e-35

>TREMBL:AF145690_1 gene: "BcDNA.LD28657"; product:
"BcDNA.LD28657";
40 Drosophila melanogaster clone LD28657 BcDNA.LD28657
(BcDNA.LD28657) mRNA,
complete cds.
Length = 637

45 HSPs:

Score = 1140 (171.0 bits), Expect = 1.3e-115, P = 1.3e-115
Identities = 230/465 (49%), Positives = 304/465 (65%)

50 Query: 61
CGRWQKRREEVNQRNVP GIDSAYLAM DTEEGVEVVWNEVQFSERK NYKLQEEKVR AVFDN 120
CGRW KRREEV+QR+VPGID +LAM DTEEGVEVVWNEVQ++ + K
QEEK+R VFDN
Sbjct: 102
55 CGRWLKRREEVDQRDVP GIDCVHLAM DTEEGVEVVWNEVQYASLQELKSQEEKMRQVFDN 161

Query: 121 LIQLEHLNIVKFHKYWADIKE-
NKARVIFITEYMSSGSLKQFLKKT KKNHKT MNEKA WK 179

```

L+QL+H NIVKFH+YW D ++ + RV+FITEYMSSGSLKQFLK+TK+N K
 + ++W+R
 Sbjct: 162
 LLQLDHQNIIVKFHRYWTDTDQAERPRVVFITEYMSSGSLKQFLKRTKRNAKRLPLESWRR 221
 5 Query: 180
 WCTQILSALSYLHSCDPPIIHGNLTCDTIFIQHNGLIKIGSVAPDTINNHVKTCEEQKN 239
 WCTQILSALSYLHSC PPIIHGNLTCD+IFIQHNGL+KIGSV PD ++ V+
 RE ++
 10 Sbjct: 222
 WCTQILSALSYLHSCSPPIIHGNLTCD SIFIQHNGLVKIGSVVPDAVHYSVRRGRERERE 281
 Query: 240 ----LHFF-APEYGEVTNVTTAVDIYSFGMCALEMAVLEIQ-
 GNGESSYVPQEAISSAIQ 293
 15 H+F APEYG +T A+DIY+FGMCALEMA LEIQ N ES+ +
 +E I I
 Sbjct: 282
 RERGAHYFQAPEYGAADQLTAALDIYAFGMCALEMAALEIQPSNSESTAINETIQRITF 341
 20 Query: 294 LLEDPLQREFIQKCLQSEPARRPTARELLFHPALFEVPSLKLLAAHCIV---
 GHQHMPE 350
 LE+ LQR+ I+KCL +P RP+A +LLFHP LFEV SLKLL AHC+V
 ++ M E
 Sbjct: 342
 25 SLENDLQRDLIRKCLNPQPDQPSANDLLFHPLLFEVHSLKLLTAHCLVFSPANRTMFSE 401
 Query: 351 NALEEITKNM-
 DTSAVLAEIPAGPGREPVTLYSQSPALELDKFLVDVRNGIYPLTAFGL 409
 A + + + V+A++ G+E L S A +L+KF+EDV+
 30 G+YPL +
 Sbjct: 402
 TAFDGLMQRYYQPDVVMAQLRLAGGQERQYRLADVSGADKLEKFVEDVKYGVYPLITYS- 460
 Query: 410
 35 PRXXXXXXXXXXXXXXXXXXXXXXXXXAEVETRKVVLMQCNIESVEEGVXXXXXXXXXXXXX 469
 + + E+R++V M C+++ E+
 Sbjct: 461
 GK KPPNFRSRAASPERADSVKSATPEPVDTESRRIVNMMCSVKIKEDSNDITMTILLRMD 520
 40 Query: 470 XXXXXXSCDLM PNENIPELAAELVQLGFISEADQSR L TS LLEETL 515
 +C + N+ +L +ELV+LGF+ DQ ++ LLEETL
 Sbjct: 521 DKMNRQLTCQVNENDTAADLTSELVRLGFVHLDDQDKIQVLEETL 566

45 Pedant information for DKFZphfbr2_78d18, frame 1

Report for DKFZphfbr2_78d18.1

50 [LENGTH] 564
 [MW] 62464.87
 [pI] 5.10
 [HOMOL] TREMBL:AF145690_1 gene: "BcDNA-LD28657"; product:
 55 "BcDNA-LD28657"; Drosophila melanogaster clone LD28657
 BcDNA-LD28657 (BcDNA-LD28657) mRNA, complete cds. le-123
 [FUNCAT] 03-22 cell cycle control and mitosis [S. cerevisiae,
 YJL095w] le-15

- [FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae,
 YJL095w] 6e-15
 [FUNCAT] 11.01 stress response [S. cerevisiae, YJL095w] 6e-15
 [FUNCAT] 03.01 cell growth [S. cerevisiae, YJL095w] 6e-15
 5 [FUNCAT] 10.02.11 key kinases [S. cerevisiae, YJL095w] 6e-15
 [FUNCAT] 03.04 budding, cell polarity and filament formation
 [S. cerevisiae, YJL095w] 6e-15
 [FUNCAT] 98 classification not yet clear-cut [S. cerevisiae,
 YLR096w] 2e-09
 10 [FUNCAT] 30.02 organization of plasma membrane [S. cerevisiae,
 YLR096w] 2e-09
 [FUNCAT] 10.03.11 key kinases [S. cerevisiae, YNR031c] 3e-09
 [FUNCAT] 09.01 biogenesis of cell wall [S. cerevisiae,
 YNR031c] 3e-09
 15 [FUNCAT] 03.07 pheromone response, mating-type determination,
 sex-specific proteins [S. cerevisiae, YLR362w] 4e-08
 [FUNCAT] 10.05.11 key kinases [S. cerevisiae, YLR362w] 4e-08
 [FUNCAT] 10.04.11 key kinases [S. cerevisiae, YLR362w] 4e-08
 [FUNCAT] 11.04 dna repair (direct repair, base excision repair
 20 and nucleotide excision repair) [S. cerevisiae, YPL153c] 1e-07
 [FUNCAT] 03.19 recombination and dna repair [S. cerevisiae,
 YPL153c] 1e-07
 [FUNCAT] 03.22.01 cell cycle check point proteins [S.
 cerevisiae, YPL153c] 1e-07
 25 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YPL153c]
 1e-07
 [FUNCAT] 03.25 cytokinesis [S. cerevisiae, YDR507c] 1e-07
 [FUNCAT] 10.99 other signal-transduction activities [S.
 cerevisiae, YPL153c] 1e-07
 30 [FUNCAT] 03.13 meiosis [S. cerevisiae, YDR523c] 3e-07
 [FUNCAT] 03.10 sporulation and germination [S. cerevisiae,
 YDR523c] 3e-07
 [FUNCAT] 03.16 dna synthesis and replication [S. cerevisiae,
 YMR001c] 2e-06
 35 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YDR490c]
 3e-05
 [FUNCAT] 05.07 translational control [S. cerevisiae, YDR283c]
 1e-04
 [FUNCAT] 01.05.04 regulation of carbohydrate utilization [S.
 40 cerevisiae, YDR477w] 1e-04
 [BLOCKS] PF00637A
 [BLOCKS] BP03191J
 [BLOCKS] PF01317B
 [SCOP] d1ir3a_ 5.1.1.2.6 insulin receptor Complex
 45 (transferase/substrate) 2e-53
 [SCOP] d1phk_ 5.1.1.1.6 gamma-subunit of glycogen
 phosphorylase kinase 3e-68
 [SCOP] d1fgkb_ 5.1.1.2.5 Fibroblast growth factor
 receptor 1 [human (Homo)] 1e-55
 50 [SCOP] d1abo_ 5.1.1.1.14 Protein kinase CK2, alpha
 subunit [Maize (Zea)] 2e-55
 [SCOP] d3lck_ 5.1.1.2.2 Lymphocyte kinase (lck) [Human
 (Homo sapiens)] 7e-54
 [SCOP] d2erk_ 5.1.1.1.11 MAP kinase Erk2 [Rat (Rattus
 55 norvegicus)] 9e-71
 [SCOP] d1cdkb_ 5.1.1.1.2 cAMP-dependent PK, catalytic
 subunit Comple 1e-55

```

[SCOP]          dlhcl__ 5.1.1.1.1 Cyclin-dependent PK [Human
(Homo sapiens) 4e-67
[EC]            2.7.1.112 Protein-tyrosine kinase 4e-06
[EC]            2.7.1.37 Protein kinase 3e-09
5  [PIRKW]       phosphotransferase 2e-28
   [PIRKW]       nucleus 3e-06
   [PIRKW]       RNA binding 3e-10
   [PIRKW]       tandem repeat 4e-07
   [PIRKW]       cell cycle control 3e-06
10 [PIRKW]       serine/threonine-specific protein kinase 2e-13
   [PIRKW]       transmembrane protein 4e-07
   [PIRKW]       autophosphorylation 3e-10
   [PIRKW]       tyrosine-specific protein kinase 4e-06
   [PIRKW]       magnesium 4e-07
15 [PIRKW]       ATP 2e-13
   [PIRKW]       receptor 4e-07
   [PIRKW]       phosphoprotein 2e-13
   [PIRKW]       apoptosis 3e-06
   [PIRKW]       glycoprotein 4e-07
20 [PIRKW]       protein kinase 2e-28
   [PIRKW]       signal transduction 2e-08
   [PIRKW]       cell division 1e-11
   [PIRKW]       calmodulin binding 3e-06
   [SUPFAM]      protein kinase byr2 1e-06
25 [SUPFAM]      unassigned Ser/Thr or Tyr-specific protein kinases 2e-
   13
   [SUPFAM]      leucine-rich alpha-2-glycoprotein repeat homology 4e-07

   [SUPFAM]      double-stranded RNA-binding repeat homology 3e-10
30 [SUPFAM]      SAM homology 1e-06
   [SUPFAM]      death-associated protein kinase 3e-06
   [SUPFAM]      ankyrin repeat homology 3e-06
   [SUPFAM]      protein kinase homology 2e-28
   [SUPFAM]      kinase-related transforming protein 2e-06
35 [SUPFAM]      protein kinase SPK1 3e-06
   [SUPFAM]      protein kinase Xa21 4e-07
   [SUPFAM]      protein kinase TIK 3e-10
   [SUPFAM]      kinase interaction domain homology 3e-06
   [PFAM]        Eukaryotic protein kinase domain
40 [KW]          All_Alpha
   [KW]          3D
   [KW]          LOW_COMPLEXITY      16.49 %

45 SEQ  IRGPGTRAGAEAAAGRGVGSAGLSVPSSMSEGESQTVLSSGSDPKVESSSSAPGLTSVS
   SEG  .....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
   Ikoba .....

50 SEQ  PPVTSTTSAASPEEEEESEDESEILEESPCGRWQKRREEVNQARNVPGIDSAYLAMDTEEG
   SEG  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
   Ikoba .....

55 SEQ  VEVVWNEVQFSEKKNYKLQEEKVRAVFDNLIQLEHLNIVKFHKYWADIKENKARVIFITE
   SEG  .....
   Ikoba .....CHHHHHHHHHHHHHHHHTTTBTTBCCEE----
   EEEETTTEEEEEEC

```



```

5  SEQ  YMSSGSLKQFLKKTCKNHKTMNEKA WKRWCTQILSALSYLHSCDPPIIHGNLTCDTIFIQ
   SEG  .....
   Jkoba  CCCCCEH--HHHHCTTTTC-CCHHHHHHHHHHHHHHHHHHH--
   HHCEETTTTTTTEETT

10  SEQ  HNGLIKIGSVAPDTINNHVKTCREEQKNLHFFAPEYGEVTNVTTAVDIYSFGMCALEMAV
   SEG  .....
   Jkoba  TTCCEEECCCTTTTEECTTTTEEEEEETTTGGGCCCHHHHHCCCBCHHHHHHHHHHHHHHHHC

15  SEQ  LEIQGN GESSYVPQEAISSAIQLLEDPLQREFIQKCLQSEPARRPTARELLFHPALFEVP
   SEG  .....
   Jkoba  CCTTTTCCCHHHHHHHHHHCCCCTTTHHHHHHHHHHTTTTTGGGCCCHHHHHHTTTT....

20  SEQ  SLKLLAAHCIVGHQHMIPENALEEITKNM DTS AVLAEIPAGPGREPVQTLYSQSPALELD
   SEG  .....
   Jkoba  .....

25  SEQ  KFLEDVRNGIYPLTAFGLPRPQAPQAEVTS PVVPPSVKTP TPEPAEVETRKVVL MQCNI
   SEG  .....
   Jkoba  .....

30  SEQ  ESVEEGVKHHLTLLLKLEDKLNRLHSCDLMPNENIPELAAELVQLGFISEADQSRLTSL
   SEG  .....
   Jkoba  .....

35  SEQ  EETLNKFN FARNSTLNSAAVTVSS
   SEG  .....
   Jkoba  .....

```

(No Prosite data available for DKFZphfbr2_78d18.1)

40 Pfam for DKFZphfbr2_78d18.1

HMM_NAME Eukaryotic protein kinase domain

```

45  HMM
   *rLnHPNIIRFYDwFed...ddDHIYMIMEYMeGGDLFDYIrrng....p
                                     +L H NI++F ++ D + ++ +I+EYM G+L +++++ +
   Query      152
   QLEHLNIVKFHKYQADIKENKARVIFITEYMSSGSLKQFLKKTCKNHKT  200

50  HMM
   MsEweIrfIMyQILrGMeYLHSMg..IIHRDLKPENILIDeNgqIKIcDF
                                     M+E+ +++ +QIL++++YLHS IIH L + I+I +NG
   IKI+
   Query      201
   MNEKA WKRWCTQILSALSYLHSCDPPIIHGNLTCDTIFIQHNGLIKIGSV  250

```

HMM

GLARqMnnYerMttfCGTPWYMMAPEVImgnyYttkVDMWSFGCILWEM

++ N+ + + APE + ++ TT+VD++SFG+

EM

5 Query 251 APDTINNHVKTCREEQKNLHFF-APEY-
 GEVTNVTTAVDIYSFGMCALEM 298

HMM

MTGepPFyddnMemImrIiqrfrppfWpnCSeElyDFMrwCUnyDPekRP

+ ++ + N E + ++ + ++ + + ++F+ +C++

P++RP

10 Query 299 A--VLEIQ-
 GNGESSYVPQEAISSAIQLLEDPLQREFIQKCLQSEPARRP 345

HMM

TFrQILnHPWF*

T+R++L HP +

15 Query 346 TARELLFHPAL 356

DKFZphfbr2_78d4

group: transmembrane protein

DKFZphfbr2_78d4 encodes a novel 188 amino acid protein without similarity to known proteins.

The novel protein contains 1 transmembrane region and a Cytochrome c family heme-binding site.

No informative BLAST results; No predictive prosite, pfam or SCOP motif.

The new protein can find application in studying the expression profile of brain-specific genes and as a new marker for amygdala cells.

weak similarity to hypothetical protein of *Arabidopsis thaliana*

perhaps complete cds.

Pedant: TRANSMEMBRANE 1

Sequenced by MediGenomix

Locus: unknown

Insert length: 1547 bp

Poly A stretch at pos. 1527, polyadenylation signal at pos. 1508

```
1 TTGCCGCCGC CGCCACCCCC GCCCAGGATG GCGGAAGTGG AGGCGCCGAC
51 GCGGGCCGAG ACGGACATGA AGCAATATCA AGGCTCCGGC GCGCTCGCCA
101 TGGATGTGGA ACGGAGTCGC TTCCCTACT GCGTGGTGTG GACGCCCATC
35 151 CCGGTGCTCA CGTGGTTTTT CCCCATCATC GGCCACATGG GCATCTGCAC
201 ATCCACAGGA GTCATTGCGG ACTTCGCGGG CCCCTACTTT GTCTCAGAGG
251 ACAACATGGC CTTTGGAAG CTTGCCAAGT ACTGGAAGTT GGACCCTGCT
301 CAGGTCTATG CTAGCGGGCC CAACGCATGG GACACGGCTG TGCACGACGC
351 CTCTGAGGAG TACAAGCACC GCATGCACAA TCTCTGCTGT GACAACTGCC
40 401 ACTCGCACGT GGCATTGGCC CTGAATCTGA TGCCTACAA CAACAGCACC
451 AACTGGAATA TGGTGACGCT CTGCTTCTTC TGCCTGCTCT ACGGGAAGTA
501 CGTCAGCGTT GGGGCCTTCG TGAAGACCTG GCTGCCCTTC ATCCTTCTCC
551 TGGGCATCAT CCTCACCGTC AGCCTGGTCT TTAACCTCCG GTGATGGCTG
601 CTCGGTGGCC CCACACCCAC CAGGGTCCCG AGGAAACAGC CGCCATCCCT
45 651 TTTGGTTCCA GATTTTTTTC TCCTCACCCC AAAAGGCAGG GTTGGGCCTG
701 CTGTTGTGGA CCGGGGGTCC GGGCTGGCAG GATGGAAGGA CTGAGGACCA
751 GCATGAAGTG GGGGTTTGTG GTCTCCCTGC CTCTCAGAAG CACCCTGTCC
801 CCTCCTCCCC AGGCCTGTGA CTCCGGCCCT GGAAGCCCTT TTGTTCTTCT
851 GTTGAAAGGC TTTGGCTTCC CTCTGTAGAG CTGCTCCCGC CACCACCTGC
50 901 TGGGGTCCTG CCTCAGCCCA GTGCCCAGTA TGGGGAGAGG AGGACATTTG
951 GGCTCACCTG TCAAGGTGGC CCTGGGACCA GAGCTGGTCC CAGCATGGGG
1001 TGCACCGGGT ACACTTAACG TGTCTCTATA AGCCAAGTTG CTTCAGGACC
1051 TTCACCACTG GCCTCTAGAA TGGTCCAGAG GGGCTGGCTG GGTCCCTTTG
1101 TCAGACTCCT GCGGGCAGCT GCCCTGGGG ACATGTGTGC CCATCTGGCA
55 1151 TCCTCCAGCC CGTCCAGTCC GCTCTTCACT GTTCCACGGC CTCCAGTGC
1201 TCCCAGCAT TGGACCCATC TCCCCTGCA GTTTGAGGCC AGAGAGGTGA
1251 GTGGACCTGA CAAGTGCCAG AGTAACCGTG TAGACAGAGC AGTGTAGACA
1301 GCGCTCAGCC CCAGCCCCAG GTGTGGACCT CATGCTGGTG ATGGCTCCCC
```



1351 TGGGTGGCCT GCCAGCACAG CCAGTGCCAT CAGGGAGCTG AAGGGGCTGT
1401 CCCCCACCTA ACTCCAGCTC CCCCTTCACG TTGTCACCAA GGCCCTGTGC
1451 CGCCCGCCTC GCCCCCCTGC TCTGTGGATT CCTTTGGGAA GGGCTCCCTG
1501 GGCAGGACAA TAAAGAGTTT TGA CTCCAAA AAAAAAAAAA AAAAAA

5

BLAST Results

10 Entry T02616 from database PIR:
hypothetical protein T19L18.12 - Arabidopsis thaliana
Score = 229, P = 1.3e-17, identities = 57/161, positives =
78/161,
frame +1

15

Medline entries

20

No Medline entry

25

Peptide information for frame 1

ORF from 28 bp to 591 bp; peptide length: 188
Category: similarity to unknown protein
30 Classification: no clue
Prositate motifs: CYTOCHROME_C (121-119)

35

1 MAEVEAPTAA ETDMKQYQGS GGVAMDVERS RFPYCVVWTP IPVLTWFFPI
51 IGHMGICTST GVIRDFAGPY FVSEDNMAFG KPAKYWKLD P AQVYASGPNA
101 WDTAVHDASE EYKHRMHNLC CDNCHSHVAL ALNLMRYNNS TNWNMVTLCF
151 FCLLYGKYVS VGAFVKTWLP FILLGII LT VSLVFNL R

40

BLASTP hits

No BLASTP hits available

45

Alert BLASTP hits for DKFZphfbr2_78d4, frame 1

PIR:T02616 hypothetical protein T19L18.12 - Arabidopsis thaliana,
N =
2, Score = 226, P = 4.5e-21

50

>PIR:T02616 hypothetical protein T19L18.12 - Arabidopsis thaliana
Length = 267

55

HSPs:

Score = 226 (33.9 bits), Expect = 4.5e-21, Sum P(2) = 4.5e-21
Identities = 52/132 (39%), Positives = 71/132 (53%)

Query: 25
MDVERSRFPYCVVWTPIPVLTWFFPIIGHMGICTSTGVIRDFAGPYFVSEDNMAFGKPAK 84
+D ++S+FP C+VWTP+PV++W P IGH+G+C GVI DFAG F++ D+

5 AFG PA+
Sbjct: 61
IDTKKSKFPCCIVWTPLPVVSWLAPFIGHIGLCREDGVILDFAGSNFINVDDFAFGPPAR 120

Query: 85 YWKLDPAQVYASGPNWDTAVHDAEEYKHRMHNLC--
10 CDNCHSHVALALNLMRYNNST- 141
Y +LD + PN H +KH DN S +

YN T
Sbjct: 121 YLQLDRTKCCLP-PNMGG---
HTCKYGFKHTDFGTARTWDNALSSSTRSFEHKTYNIFTC 176

15 Query: 142 NWN-MVTLCFFCLLYG 156
N + V C L YG
Sbjct: 177 NCHSFVANCLNRLCYG 192

20 Score = 157 (23.6 bits), Expect = 1.8e-13, Sum P(2) = 1.8e-13
Identities = 27/81 (33%), Positives = 50/81 (61%)

Query: 101
WDTAVHDAEEYKHRMHNLC DNCHSHVALALNLMRYNNSTNWNMVTLCFFCLLYGKYVS 160
25 WD A+ ++ ++H+ +N+ NCHS VA LN + Y S WNMV +

++ GK+++
Sbjct: 155
WDNALSSSTRSFEHKTYNIFTCNCHSFVANCLNRLCYGGSMEWNMVNVAILLMIKGKWIN 214

30 Query: 161 VGAFVKTWLPFILL--LGIIL 179
+ V+++LP ++ LG++L
Sbjct: 215 GSSVRSFLPCAVVTSLGVL 235

Score = 36 (5.4 bits), Expect = 4.5e-21, Sum P(2) = 4.5e-21
35 Identities = 7/21 (33%), Positives = 14/21 (66%)

Query: 10 AETDMKQYQSGGGVAMDVERS 30
++ ++K +G G MD++RS
40 Sbjct: 12 SDRNLKMSRGRGVPMMDLKRS 32

Pedant information for DKFZphfbr2_78d4, frame 1

45 Report for DKFZphfbr2_78d4.1

50 [LENGTH] 188
[MW] 21178.66
[pI] 6.27
[HOMOL] PIR:T02616 hypothetical protein T19L18.12 -
Arabidopsis thaliana 7e-32
[PROSITE] CYTOCHROME_C 1
[KW] TRANSMEMBRANE 1

55 SEQ MAEVEAPTAAETDMKQYQSGGGVAMDVERSRFPYCVVWTPIPVLTWFFPIIGHMGICTST
PRD cccccchhhhhhhhhhhcc

DKFZphfbr2_78e18

5 group: brain derived

DKFZphfbr2_78e18 encodes a novel 307 amino acid protein without similarity to known proteins.

10 The mRNA is differentially polyadenylated.
No informative BLAST results; No predictive prosite, pfam or SCOP motif.

15 The new protein can find application in studying the expression profile of brain-specific genes.

similarity to hypothetical protein of Arabidopsis thaliana

20 differential polyadenylation
> 7 exons
complete on human genomic clone 451B21ap.
perhaps complete cds.

25 Sequenced by MediGenomix

Locus: /map="144.50 cR from top of Chr6 linkage group"

Insert length: 3096 bp

30 Poly A stretch at pos. 3075, polyadenylation signal at pos. 3047

```

      1 TGGTGAGTTC GGAGTAGAGA TGGCCGCGCT TGCACCGCTG CCCCCGCTCC
      51 CCGCACAGCT CAAGAGCATA CAGCATCATC TGAGGACGGC TCAGGAGCAT
35    101 GACAAGCGAG ACCCTGTGGT GGCTTATTAC TGTCGTTTAT ACGCAATGCA
      151 GACTGGAATG AAGATCGATA GTAAACTCC TGAATGTCGC AAATTTTTAT
      201 CAAAGTTAAT GGATCAGTTA GAAGCTCTAA AGAAGCAGTT GGGTGATAAT
      251 GAAGCTATTA CTCAAGAAAT AGTGGGCTGT GCCCATTGGG AGAATTATGC
      301 TTTGAAAATG TTTTGTATG CAGACAATGA AGATCGTGCT GGACGATTTT
40    351 ACAAAAACAT GATCAAGTCC TTCTATACTG CAAGTCTTTT GATAGATGTC
      401 ATAACAGTAT TTGGAGAACT CACTGATGAA AATGTGAAAC ACAGGAAGTA
      451 TGCCAGATGG AAGGCAACAT ACATCCATAA TTGTTTAAAG AATGGGGAGA
      501 CTCCTCAAGC AGGCCCTGTT GGAATTGAAG AAGATAATGA TATTGAAGAA
      551 AATGAAGATG CTGGAGCAGC CTCTCTGCCC ACTCAGCCAA CTCAGCCATC
45    601 ATCATCTTCA ACTTATGACC CAAGCAACAT GCCATCAGGC AACTATACTG
      651 GAATACAGAT TCCTCCGGGT GCACACGCTC CAGCTAATAC ACCAGCAGAA
      701 GTGCCTCACA GCACAGGTGT AGCAAGTAAT ACTATCCAAC CTACTCCACA
      751 GACTATACCT GCCATTGATC CCGCACTTTT CAATACAATT TCCCAGGGGG
      801 ATGTTCTGCT AACCCAGAA GACTTTGCTA GAGCTCAGAA GTACTGCAAA
50    851 TATGCTGGCA GTGCTTTGCA GTATGAAGAT GTAAGCACTG CTGTCCAGAA
      901 TCTACAAAAG GCTCTCAAGT TACTGACGAC AGGCAGAGAA TGAAGCCTTT
      951 GTATGACAGA CCCATGTATT TTTGGCATGA GGAACATAACA GTCCATTACT
100   1001 CTATCTTCAG CCTATCAGGA TCACAGTTT AAGGAAGACT TGGTTTTGTT
105   1051 GAATATGACA ATGAAATCTG TGTGTATCAG ATTTTATTG AAGCATTTCAT
55   1101 CAGCAGCCTC AACCAGTTT CATTGTCCAT TTACTAGATT CAATCGTCTC
115   1151 TGAGTATATA GGGCTGATGT TAGCAAGACC CTAAAAATGT CCATTGAACC
120   1201 CTGCTTCAAA AAATGAAAC ACACCTCTAT AAAATGTGTA CTGGGAATAA
125   1251 GCTTTGTATT TACATACATT AGGGGAATTT TTTAAAATCT GTAATGTTTG

```


1301 GACAAACAGA TGATATTACT TTGCTATAAA ATTATAAATG TAACCTTTTAA
1351 TAAAGATAGC CAGAATATTC TAAATTAGAA ATTACGTTTT TGTTTCCCTC
1401 AAGACATAAA ACAAATATAA ACATTCTAAA CTGCTGGATG AATCTGAAAA
1451 GACATTAAAGT TCAAATTTTA ATTTATTCTC ATATTAAATA TAACCTCCATT
5 1501 AAAAGTTTAA AATTTTCATGG GAGAAAATAT AATAAGGTAA AGAGGTAGAA
1551 TCACTTTCAG ACTTAAGAAT AATGTTGATT TCCCAAGTGC TTTACCTTAT
1601 CTGTTAAAGC GTAAGATGAA TTGGTATTTG CTTCATAGGC AGTTTGGACTG
1651 CATGTATTAG AGAATGAAAA GAAGATATTT GTAGTAATGC CTGGAAACTT
1701 GGTGCTTTAA ATTAAGGTAC TCCTCTGCTG CTGTAGAATG GATTCCACAC
10 1751 AGTGGATAGC TATGGGTGAT TCAGAATATT ATGTTTAGAT TCCCATTGTG
1801 TAAGTTTATA AGTTTTGTGG GGAATTATGA ACTTACTGTG TACTACCTGC
1851 ATTTGTGCTG TGTGAAAAAT AAATACAAGG ATTCGTTTAG CTAATTCAAC
1901 TTAATAAAA GACAAATGTC TGTTTTTATT TGCCTGCTAG GATTGTCTTT
1951 TTTAAAAGTC ATTTTTATTT ATAGGAATAT GGGTGTCTTCT ATAGGAAGAA
15 2001 ACAGGTTTTT TGTTTTTTGT TTTTAAAGAT AAATTTGACA AAGTTAACTG
2051 AAATTTATCT GGTCCATTTT ATTCATGCTA CTAAGATGGG AATCTTTAAA
2101 CACAAGGGTC AGCAAGCTTT GGCCCATGGA TTGGCCACCT GTTACGTAAA
2151 TAAAGTTTCT TTGAAACAAG CCTACACTCA TTCATTTATG TTTTGTCTGT
2201 GGTGCTTTT CACAACCTGCA GAGTTGTATG GCTTGCAAGT CTA AAAACAT
20 2251 TTAATAATTTG GCCCTCTAAG AAAAAAGTTAA GACACCTAGT CTAATGGCCT
2301 TTTGGGAAAA AACAAATCAC TAACTCATAA TCATTTATAT CCATTATTTT
2351 CTGCATAAAT GTAATGCTAT TGTACAGGGT TTGGTAGAAT AAATATTCAG
2401 ACTGACTAAA CTGTTCTAAA TTCTCACAAA AAAGTCCCCA AACAACATGC
2451 CTCCTAAAAA ACATTTTCTT ATCTTTTACA AGAGGTATGA ACATTTGTAG
25 2501 GGTTCCACAT TTGCATCTAG AAATCCAATG CTCTTTAGAA TGTTATTACG
2551 AATAGAAAAGA TGGCCAGGAT GACCTTTAGT GTTACATGAT GTTCAGCAAA
2601 TTTTAATTCA AACCTTGATA TGCCTGGACA CTGAAAAGTA AACGCATCAC
2651 CTCCTATTTT ATACCCTACC TTCTGGTTCC CAATTGGGAG AGCACATAGA
2701 GGGGAAGGAGA CAATATAGAA ACTACGGAGT CCGCTGGTAG TGGGCTGCAT
30 2751 GGTGTGACAG AGCCCTTCTC TGTA AAAATGG AAATGACACC ACTAGCCATC
2801 TCAATAGTTA CAAGAATTAA AAGAGATACA GTACCTGAAG TGCTTAGCGC
2851 ATGGTAGCAT TTCATAAATG TTTAGTGTCA ATACTAATGC TCTAATAATG
2901 TAAATTGTTA ATAATTTATT TCCCTAATAT CAGGAAATCC CAGTTGTCTA
2951 TGTGGCCGAG TGCTTAAAAA CGCCTTCTTG CATGAGGGGA TTGAACTATA
35 3001 CAATGTTTGT TAACTTTGTA TTTGTATTTT TTCCTATAAA ATCTTAAAT
3051 AAAATTAGGA GATGTGTTCT GATGTAAAAA AAAAAAAA AAAAAA

BLAST Results

Entry HS451B21 from database EMBL:

Human DNA sequence *** SEQUENCING IN PROGRESS *** from clone
451B21

45 Score = 11219, P = 0.0e+00, identities = 2287/2343

Medline entries

No Medline entry

Peptide information for frame 2

ORF from 20 bp to 940 bp; peptide length: 307
 Category: similarity to unknown protein
 Classification: no clue

```

5      1 MAALAPLPPL PAQLKSIQHH LRTAQEHDKR DPVVAYYCRL YAMQTGMKID
      51 SKTPECRKFL SKLMDQLEAL KKQLGDNEAI TQEI VGCAHL ENYALKMFLY
     101 ADNEDRAGR F HKNMIKSFYT ASLLIDVITV FGELTDENVK HRKYARWKAT
     151 YIHNCLKNGE TPQAGPVGIE EDNDIEENED AGAASLPTQP TQSSSSSTYD
     201 PSNMPSGNYT GIQIPPGAHA PANTPAEVPH STGVASNTIQ PTPQTIPAI
10    251 PALFNTISQG DVRLTPEDFA RAQKYCKYAG SALQYEDVST AVQNLQKALK
     301 LLTTGRE
  
```

15 BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphfbr2_78e18, frame 2

20 No Alert BLASTP hits found

Pedant information for DKFZphfbr2_78e18, frame 2

25 Report for DKFZphfbr2_78e18.2

```

30  [LENGTH] 313
    [MW]     34463.95
    [pI]     5.64
    [HOMOL]  PIR:T04798 hypothetical protein F10M23.90 -
    Arabidopsis thaliana 3e-22
    [KW]     All_Alpha
35  [KW]     LOW_COMPLEXITY 16.61 %
  
```

```

40  SEQ  GEFGVEMAALAPLPPLPAQLKSIQHHLRTAQEHDKRDPVVAYYCRLYAMQTGMKIDSKTP
     SEG  .....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
     PRD  ccchhhhhhheeeccccchhhhhhhhhhhhhhhhhhhhhccceehhhhhhhhhhhccccc

45  SEQ  ECRKFLSKLMDQLEALKKQLGDNEAITQEI VGCAHLENYALKMFLYADNEDRAGR FHKM
     SEG  .....
     PRD  chhhhhhhhhhhhhhhhhhhhhccccccccchhhhhhhhhhhhhhhhhhhhhccccccccchh

50  SEQ  IKSFYTASLLIDVITVFGELTDENVKHRKYARWKATYIHNCLKNGETPQAGPVGIEEDND
     SEG  .....xxxxxxxx
     PRD  hhhhhhhhhhhhhhhhhhhhhccccccccchhhhhhhhhhhhhhhhhhhhhcccccccccccc

55  SEQ  IEENEDAGAASLPTQPTQPSSSSTYDPSNMPSGNYTGIQIPGAHAPANTPAEVPHSTGV
     SEG  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
     PRD  ccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

     SEQ  ASNTIQPTPQTIPAI DPALFNTISQGDVRLTPEDFARAQKYCKYAGSALQYEDVSTAVQN
     SEG  .....
     PRD  cccccccccccccccccccccccccccccccccccccchhhhhhhhhhhhhhhccceecchhhhhh

     SEQ  LQKALKLLTTGRE
  
```

SEG
PRD hhhhhhhhcccc

5 (No Prosite data available for DKFZphfbr2_78e18.2)

(No Pfam data available for DKFZphfbr2_78e18.2)

DKFZphfbr2_78i21

5 group: metabolism

DKFZphfbr2_78i21 encodes a novel 477 amino acid protein with similarity to beta-aspartate methyltransferases.

10 The L-isoaspartyl methyltransferase (Pimt), as an example, is a highly conserved enzyme utilising S-adenosylmethionine (AdoMet) to methylate aspartate residues of proteins damaged by age-related isomerisation and deamidation.

15 The new protein can find application in diagnosis/modulation of protein damage and age-related degenerative processes.

unknown protein

20

weak similarity to beta-aspartate methyltransferase pimT of *Mycobacterium leprae* perhaps complete cds.

25 Sequenced by MediGenomix

Locus: unknown

Insert length: 1842 bp

30 Poly A stretch at pos. 1819, polyadenylation signal at pos. 1802

```
1 CCTTCGCGAA ACACTATGCT AATGGCATGG TGCCGCGGTC CTGTCTTGCT
51 GTGCCTGCGG CAGGGGGCTCG GAACCAATTC ATTCCTGCAC GGCCTGGGGC
35 101 AGGAGCCCTT CGAGGGAGCT CGGTCACTGT GTTGCAAGGC CTCGCCTAGA
151 GACCTGCGAG ATGGAGAAAG AGAGCACGAG GCGGCACAAA GGAAAGCCCC
201 AGGAGCAGAG TCTTGCCCAT CTCTCCCTCT GAGCATCTCG GACATTGGGA
251 CTGGATGTCT TTCGTCACTG GAAAACCTCA GACTGCCGAC GCTGCGGGAA
301 GAGTCATCCC CTCGAGAGCT CGAGGACTCG AGCGGAGACC AGGGCCGGTG
40 351 CGGTCCCACA CACCAGGGAT CCGAGGATCC TTCGATGCTC TCGCAGGCCC
401 AGTCCGCTAC CGAGGTCGAA GAGCGTCACG TCTCCCCTTC TTGTTCAACT
451 TCCAGAGAGA GACCCTTTCA GGCTGGGGAA CTGATTTTAG CTGAGACTGG
501 GGAGGGAGAA ACAAATTTA AGAAATTATT TAGGTTGAAC AACTTCGGAC
551 TCTTAAATAG TAACTGGGGG GCAGTCCCGT TCGGCAAGAT CGTGGGGGAG
45 601 TTCCCCGGCC AGATACTGAG GAGTTCCTTC GGTAAGCAGT ACATGCTGAG
651 GAGGCCAGCC TTGGAAGACT ATGTAGTATT GATGAAAAGA GGGACTGCCA
701 TAACATTCCC AAAGGATATT AATATGATTC TCTCAATGAT GGATATCAAC
751 CCAGGTGATA CTGTTTTGGA AGCTGGCTCA GGCTCTGGTG GAATGAGCTT
801 ATTTTATATCC AAAGCAGTTG GATCACAAGG ACGAGTCATA AGTTTTGAGG
50 851 TACGAAAAGA CCACCATGAT CTGGCTAAGA AGAATTACAA ACACTGGCGT
901 GATTCATGGA AATTAAGTCA TGTAGAAGAG TGGCCAGACA ATGTGGATTT
951 TATTCATAAG GACATTTTCA GAGCAACCGA AGACATAAAA TCTTTAACAT
1001 TTGACGCAGT AGCTTTGGAT ATGTTAAATC CTCATGTTAC TTTGCCTGTT
1051 TTTTACCCAC ATCTTAAGCA TGGTGGTGTA TGTGCTGTAT ATGTAGTAAA
55 1101 CATCACACAG GTTATTGAAC TTTTAGATGG AATTGCGACC TGTGAACCTG
1151 CTCTTTCATG TGAAAAGATA AGCGAGGTCA TTGTCAGAGA TTGGTTGGTT
1201 TGCCTTGCAA AACAGAAAAA TGGGAATTTTA GCTCAAAAAG TAGAATCTAA
1251 AATCAACACA GATGTACAAC TAGATTCTCA AGAGAAAATT GGAGTTAAAG
```

1301 GTGAGCTGTT TCAAGAGGAT GACCATGAAG AATCGCATTG TGATTTTCCA
1351 TATGGATCAT TTCCCTATGT TGCTAGACCA GTACACTGGC AACCTGGTCA
1401 TACAGCTTTT CTTGTCAAGT TGAGGAAGGT CAAACCACAA CTAACTGAG
1451 TACTCCAGAT GACAGTAACT GACTTGAAGA TGGAAAAATA TCAAAATAGA
5 1501 ACTTTATATT GAAAATCACT GCTTCCATAG ATTGGCATTG TTAGCTATTA
1551 CTATGACTTA TATAACTTAT ACATATAATT TTGAAAATAA CAACTAAAAG
1601 ATGTATAACA TAGCAAACT GCTTAAACAT CCCATTTTGA CACTTGTCTT
1651 GCAGTTAGTT TGACATTTTG TAGTTAATGA TTCCAAATTG GTTTAGTTGG
1701 GCCATCTCAT TCTTCACTTC CTGTAAACCA CTCCATAGAT TTGTCTTTCT
10 1751 TCAAGAAATT AGTTTTCTTT CTTTATTTG ATTGATGGTC ATTGACTACT
1801 GAAATAAAAT ATGCATTTTA AGAAAAAATA AAAAAAATAA AA

BLAST Results

15

No BLAST result

20

Medline entries

No Medline entry

25

Peptide information for frame 1

30 ORF from 16 bp to 1446 bp; peptide length: 477
Category: putative protein
Classification: no clue

35 1 MLMAWCRGPV LLCLRQGLGT NSFLHGLGQE PFEGARSLCC RSSPRDLRDG
51 EREHEAAQRK APGAESCPSL PLSISDIGTG CLSSLENLRL PTLREESSPR
101 ELEDSSGDQG RCGPTHQGE DPSMLSQAQS ATEVEERHVS PSCSTSRERP
151 FQAGELILAE TGEGETKFKK LFRLNNFGLL NSNWGAVPFG KIVGKFPGQI
201 LRSSFQKQYM LRRPALEDYV VLMKRGTAIT FPKDINMILS MMDINPGDTV
251 LEAGSGSGGM SLFLSKAVGS QGRVISFEVR KDHHDLAKKN YKHWRDSWKL
40 301 SHVEEWPQNV DFIHKDISGA TEDIKSLTFD AVALDMLNPH VTLPVFYPHL
351 KHGGVCAVYV VNITQVIELL DGIRTCELAL SCEKISEVIV RDWLVLAKQ
401 KNGILAQKVE SKINTDVQLD SQEKIGVKGE LFQEDDHEES HSDFPYGSFP
451 YVARPVHWQP GHTAFLVKLR KVKPQLN

45

BLASTP hits

No BLASTP hits available

50

Alert BLASTP hits for DKFZphfbr2_78i21, frame 1

No Alert BLASTP hits found

55

Pedant information for DKFZphfbr2_78i21, frame 1

Report for DKFZphfbr2_78i21.1

5 [LENGTH] 482
 [MW] 53521.20
 [pI] 6.28
 [HOMOL] TREMBL:AF088800_2 product: "unknown"; Rhodococcus
 erythropolis ARC (arc) gene, complete cds; and unknown genes. 2e-
 23
10 [FUNCAT] r general function prediction [M. jannaschii,
 [MJ0134] 6e-10
 [FUNCAT] 05.07 translational control [S. cerevisiae, YJL125c]
 6e-04
 [BLOCKS] BLO0801E
 [BLOCKS] BLO1279A
15 [KW] Alpha_Beta
 [KW] LOW_COMPLEXITY 2.49 %

20 SEQ PSRNTMLMAWCRGPVLLCLRQGLGTNSFLHGLGQEPFEGARSLCCRSSPRDLRDGEREHE
 SEG
 PRD ccccccccccccccccchhhhhccccccccccccccccccccccccccccccccchhhhh

 SEQ AAQRKAPGAESCPSLPLSISDITGTGCLSSLENLRLPTLREESSPRELEDSSGDQGRCGPT
 SEG
25 PRD hhhhhcc

 SEQ HQGSEDPMSLSQAQSAATEVEERHVSPPSCSTSRERPFQAGELILAETGEGETKFKKLFRLN
 SEG
30 PRD ccccccccchhh

 SEQ NFGLLNSNWGAVPFGKIVGKFGQILRSSFGKQYMLRRPALEDYVVVLMKRGTAITFPKDI
 SEG
 PRD ccchhhhhhhhhhhccceeecccc

35 SEQ NMILSMMDINPGDTVLEAGSGSGGMSLFLSKAVGSQGRVISFEVRKDHHDLAKKNYKHWR
 SEG xxxxxxxxxxxx.....
 PRD cceeccccccccccccccccccccchhh

40 SEQ DSWKLSHVEEWPDNVDFIHKDISGATEDIKSLTFDAVALDMLNPHVTLVPVFYPHLKHGGV
 SEG
 PRD hccchhhhhhhhhhhcccccc

 SEQ CAVYVVNITQVIELLDGIRTCELALSCEKISEVIVRDWLVLCLAKQKNGILAQKVESKINT
 SEG
45 PRD eeeeeechhh

 SEQ DVQLDSQEKIGVKGELFQEDDHEESHSDFPYGSFPYVARPVHWQPGHTAFLVKLRKVKPQ
 SEG
50 PRD cc

 SEQ LN
 SEG ..
 PRD cc

55

(No Prosite data available for DKFZphfbr2_78i21.1)

(No Pfam data available for DKFZphfbr2_78i21.1)

DKFZphmel2_12j1

5

group: melanoma derived

DKFZphmel2_12j1 encodes a novel 905 amino acid protein, which has similarity to integrin I of *Saccharomyces cerevisiae*.

10

The novel protein contains a leucin zipper.

No informative BLAST results; No predictive prosite, pfam or SCOP motif.

15

The new protein can find application in studying the expression profile of melanoma-specific genes.

weak similarity to integrin I (*Saccharomyces cerevisiae*)

20

Sequenced by EMBL

Locus: unknown

25

Insert length: 2942 bp

Poly A stretch at pos. 2926, no polyadenylation signal found

```

      1 CGAAAGCTAA AGGCCGGCGC ACGCTGGGCG GTGGTGGTCC CTAAGCCGGG
30    51 CCGCGGCCGG TGCAATGGAC TCCACTGCCT GCTTGAAGTC CTTGCTCCTG
      101 ACTGTCAGTC AGTACAAAGC CGTGAAGTCA GAGGCGAACG CCACTCAGCT
      151 TTTGCGGCAC TTGGAGGTAA TTTCTGGACA GAAACTCACA CGACTATTTA
      201 CATCAATCA GATATTAACA AGTGAATGCT TGAGTTGCCT TGTAGAGCTA
      251 CTTGAAGACC CCAACATAAG TGCTTCACTG ATCTTAAGTA TTATCGGTTT
35    301 GCTGTCTCAA CTAGCAGTAG ACATTGAAAC CAGAGATTGT CTTCAGAATA
      351 CATATAATCT GAATAGTGTG CTGGCGGGAG TGGTTTGTCT GAGCAGCCAC
      401 ACTGATTCGG TGTTTTTGCA GTGCATTCAA CTTCTACAGA AGTTAACATA
      451 TAATGTCAAA ATTTTCTATT CTGGTGCCAA TATAGATGAA TTAATTACGT
      501 TCCTGATAGA TCACATTCAA TCTTCTGAAG ATGAGTTAAA AATGCCTTGT
40    551 CTAGGATTAT TGGCAAATCT TTGTCGGCAC AATCTTTCTG TTCAAACGCA
      601 CATAAAGACA TTGAGTAATG TGAAATCTTT TTATCGAACT CTTATCACCT
      651 TGTTGGCCCA TAGTAGTTTA ACTGTGGTTG TGTTTGCACT TTCAATATTA
      701 TCCAGTTTGA CATTAAATGA AGAGGTGGGG GAAAAGCTAT TCCATGCTCG
      751 AAACATTTCAT CAGACTTTTC AACTAATATT TAATATTCTC ATAAACGGTG
45    801 ATGGCACTCT AACTAGAAAG TATTCAGTTG ACCTACTGAT GGATCTCCTT
      851 AAGAATCCTA AAATTGCTGA TTATCTCACC AGATATGAGC ACTTTTCTTC
      901 ATGTCTTCAC CAAGTATTAG GTCTTCTTAA TGGAAAGGAT CCTGATTCCCT
      951 CTTCAAAGGT TTTAGAATTA CTTCTTGCCT TCTGTTCACT GACTCAGCTG
50   1001 CGCCATATGC TCACTCAGAT GATGTTTGAA CAGTCTCCAC CTGGCAGCGC
      1051 CACTCTGGGA AGCCATACTA AATGTTTGA ACCTACTGTG GCTCTACTGC
      1101 GCTGGTTAAG CCAACCTTTG GACGGATCAG AAAACTGTTC TGTTTTAGCA
      1151 TTGGAGTTGT TCAAGGAAAT ATTTGAGGAT GTCATAGATG CTGCTAACTG
      1201 TTCCTCGGCT GATCGTTTTG TGACCCTTCT GCTGCCTACA ATCCTTGATC
      1251 AACTTCAGTT CACAGAACAA AATCTAGATG AGGCTTTAAC AAGAAAAAAT
55   1301 GTGAAAGGGA TTGCCAAGGC CATTGAAGTT TTGTTAACTC TCTGTGGAGA
      1351 TGATACATA AAAATGCATA TTGCAAAAAT CTTGACAACT GTCAAGTGTA
      1401 CCACTCTTAT AGAACACAA TTTACATATG GCAAGATTGA CCTGGGATTT
      1451 GGAACAAAGG TTGCAGATTC TGAATTATGC AAACCTTGCTG CTGATGTAAT

```



```

1501 TTTGAAAACCT CTTGATTTGA TTAACAAACT TAAACCATTTG GTTCCTGGTA
1551 TGGAAAGTAAG CTTCTACAAA ATACTTCAGG ACCCACGTTT GATTACTCCT
1601 TTGGCTTTTG CTTTAACGTC AGATAATAGA GAACAAGTAC AGTCTGGACT
1651 GAGAATATTA TTGGAGGCTG CTCCTCTGCC AGATTTTCCT GCTTTAGTAC
5 1701 TTGGAGAAAAG TATAGCAGCA AACAAATGCCT ATAGACAACA GGAAACAGAA
1751 CATATACCCA GAAAAATGCC CTGGCAATCA TCAAATCACA GTTTTCCAAC
1801 ATCAATAAAG TGTTTAACTC CTCATTTGAA AGATGGTGTT CCTGGATTGA
1851 ATATTGAAGA ATTAATAGAG AAACCTTCAGT CTGGAATGGT GGTAAAGGAT
1901 CAGATTTGTG ATGTGAGAAT ATCTGACATA ATGGATGTAT ATGAAATGAA
10 1951 ACTATCCACA TTAGCTTCCA AAGAAAGCAG GCTACAAGAT CTTTGGGAAA
2001 CAAAAGCTCT AGCCCTTGCA CAGGCTGATA GACTGATTGC TCAGCATCGC
2051 TGTCAAAGAA CTCAAGCTGA AACAGAGGCA CGGACACTTG CTAGTATGTT
2101 GAGAGAAGTT GAGAGAAAAA ATGAAGAGCT TAGTGTGTTG CTGAAGGCGC
2151 AGCAAGTTGA ATCAGAAAGA GCGCAGAGTG ATATTGAGCA TCTCTTTCAA
15 2201 CATAATAGGA AGTTAGAGTC TGTGGCTGAA GAACATGAAA TACTGACAAA
2251 ATCCTACATG GAACTTCTTC AGAGAAATGA AAGTACTGAA AAGAAGAATA
2301 AAGATTTACA GATCACATGT GATTCTCTGA ATAAACAAAT TGAGACAGTG
2351 AAAAAGTTGA ATGAGTCACT CAAGGAACAA AATGAAAAAA GTATTGCCCA
2401 ATTAATAGAG AAAGAAGAAC AGAGAAAAGA AGTACAGAAT CAGCTAGTAG
20 2451 ACAGAGAACA TAAGCTAGCA AATTTGCATC AAAAAACAAA AGTACAAGAA
2501 GAAAAGATTA AAACCTTACA AAAGGAAAGG GAAGATAAGG AAGAAACCAT
2551 TGATATCCTT AGAAAAGAAT TAAGCAGAAC AGAACAGATA AGAAAAGAGT
2601 TGAGCATTAA GGCTTCCTCC CTAGAGGTTT AAAAGGCACA ATTAGAAGGT
2651 CGTTTGGAAG AGAAAGAGTC CTTGGTGAAA CTTCAGCAAG AGGAATTGAA
25 2701 CAAACACTCC CACATGATAG CAATGATCCA CAGTTTAAGT GGTGGAAAAA
2751 TAAATCCAGA AACTGTGAAT CTCAGTATAT AGACATTATG GCATTTTGGG
2801 ATTTGTAATC TCATGATATT TTTGATGTAT TTATCTATTG GAGGGGGGGT
2851 GGGTAGGGGA GTTAATTTGT GACTTCGTAA CAATAAGAAG TTATTATCTA
2901 ATTTAGTAAA GACCCTGATC TGTTGCAAAA AAAAAAAAAA AA
30

```

BLAST Results

35 No BLAST result

Medline entries

40

96039111:

Hostetter MK, Tao NJ, Gale C, Herman DJ, McClellan M, Sharp RL,
Kendrick KE.: Antigenic and functional conservation of an
integrin

45

I-domain in

Saccharomyces cerevisiae. Biochem Mol Med 1995 Aug;55(2):122-30

99458454:

Berton G, Lowell CA.: Integrin signalling in neutrophils and
macrophages. Cell Signal 1999 Sep;11(9):621-35

50

55

Peptide information for frame 2

ORF from 65 bp to 2779 bp; peptide length: 905

Category: putative protein

Classification: Cellular transport and traffic

Prosite motifs: LEUCINE_ZIPPER (331-352)

```

5      1 MDSTACLKSL LLTVSQQYKAV KSEANATQLL RHLEVISGQK LTRLFTSNQI
      51 LTSECLSCLV ELLEDPNISA SLILSIIGLL SQAVIDIETR DCLQNTYNLN
     101 SVLAGVVCRS SHTDSVFLQC IQLLQKLTYN VKIFYSGANI DELITFLIDH
     151 IQSSEDELKM PCLGLLANLC RHNLSVQTHI KTLNVKSFY RTLITLLAHS
    10  201 SLTVVVFALS ILSSLTLNEE VGEKLFHARN IHQTFQLIFN ILINGDGTLT
     251 RKYSVDLLMD LLKNPKIADY LTRYEHFSSC LHQVLGLLNG KDPDSSSKVL
     301 ELLAFCSVT QLRHMLTQMM FEQSPPGSAT LGSHTKCLEP TVALLRWLSQ
     351 PLDGSSENCV LALELFKEIF EDVIDAANCS SADRFTVLLL PTILDQLQFT
     401 EQNLDEALTR KNVKGIKAI EVLLTLCGDD TLKMHIKIL TTVKCTTIE
    15  451 QQFTYGKIDL GFGTKVADSE LCKLAADVIL KTLDLINKLK PLVPGMEVSF
     501 YKILQDPRLI TPLAFALTSD NREQVQSGLR ILLEAAPLPD FPALVLGESI
     551 AANNAYRQAE TEHIPRKMPW QSSNHSFPTS IKCLTPHLKD GVPGLNIEEL
     601 IEKLQSGMVV KDQICDVRIS DIMDVYEMKL STLASKESSL QDLLETKALA
     651 LAQADRLIAQ HRCQRTQAET EARTLASMLR EVERKNEELS VLLKAQQVES
    20  701 ERAQSDIEHL FQHNKLESV AEEHEILTKS YMELLQRNES TEKKNKDLQI
     751 TCDLNLKQIE TVKKLNESLK EQNEKSIAQL IEKEEQRKEV QNQLVDREHK
     801 LANLHQKTKV QEEKIKTLQK EREDKEETID ILRKELSRTE QIRKELSIKA
     851 SSLEVQKAQL EGRLEEKESL VKLQEEELNK HSHMIAMIHS LSGGKINPET
     901 VNLSI

```

BLASTP hits

30 No BLASTP hits available

Alert BLASTP hits for DKFZphm12_12j1, frame 2

```

35  TREMBL:SCINTANA_1 Saccharomyces cerevisiae integrin analogue
    gene,
    complete cds., N = 1, Score = 216, P = 1.3e-13

```

```

40  >TREMBL:SCINTANA_1 Saccharomyces cerevisiae integrin analogue
    gene, complete
    cds.
        Length = 1,015

```

HSPs:

```

45  Score = 216 (32.4 bits), Expect = 1.3e-13, P = 1.3e-13
    Identities = 80/302 (26%), Positives = 155/302 (51%)

```

```

50  Query:   597 IEELIEKLQSGMVVKDQICDVRISDIM---
    DVVYEMKLSTLASKESSLQDLLETKALALAQ 653
          I L EKL++   D+  + +IS++   +  E +L+   + ++ L+

```

LET AL +

Sbjct: 275 ISLLKEKLETATTANDENVN-

KISELTKTREELEAELAAYKNLKNLETKLETSEKALKE 333

```

55  Query:   654 A---DRLIAQHRCQRTQAETEAR----TLASMLREVERKNEELSVLLKA--
    QQVESERAQ 704

```

+ + + + Q + TE + +L + L +E+++E+L+ LK
 +Q+ ++ Q
 Sbjct: 334
 VKENEEHLKEEKIQLEKEATETKQQLNSLRANLESLEKEHEDLAAQLKKYEEQIANKERQ 393
 5 Query: 705 SDIEHLFQHNRKLESVAEEHEILTKSYMEL---LQRNESTEKKNKDLQIT-
 CDSL NKQIE 760
 + E + Q N ++ S +E+E + K EL ++ +ST ++ +L+ +
 D+LN QI+
 10 Sbjct: 394 YN-
 EEISQLNDEITSTQQENESIKKKNDELEGEVKAMKSTSEEQSNLKKSEIDALNLQIK 452
 Query: 761
 15 TVKKL NESLKEQNEKSIAQLIEKEEQRKEVQNLVDREHKL ANLHQKTKVQEEKIKT--- 817
 +KK NE+ + +SI + + + KE+Q++ +E +++ L K.K
 E+K
 Sbjct: 453
 ELKKKNETNEASLLESISIESETVKIKELQDECNFKEKEVSELEDK LKASEDKNSKYLE 512
 20 Query: 818 LQKEREDKEETIDI----LRKELSRTEQIRKELSIKASSLE-
 VQKAQLEGRLEEKESLVK 872
 LQKE E +E +D L+ +L + + K S L ++K E R
 +E L K
 Sbjct: 513
 25 LQKESEKIKEELDAKTTELKIQLEKVTNLSKAKEKSESELSRLKKTSSSEERKNAAEQLEK 572
 Query: 873 LQQE 876
 L+ E
 Sbjct: 573 LKNE 576
 30 Score = 186 (27.9 bits), Expect = 2.0e-10, P = 2.0e-10
 Identities = 82/301 (27%), Positives = 155/301 (51%)
 Query: 598 EELIEKLQSGMVVKDQICDVRISDIMDVYEMKLSTLASKESR---LQD-
 35 LLET KALALAQ 653
 +ELI +LQ+ +K + D S++ V L K++ LQD +L
 K
 Sbjct: 611 DELI-
 RLQENELKAKEIDNTRSELEKVSLSNDELLEEKQNTIKSLQDEILSYKDKITRN 669
 40 Query: 654
 ADRLIAQHRCQRTQAETEARTLASMLREVERKNEELSVLLKAQQVESERAQSDIEHLFQH 713
 ++L++ R + E+ L LR + ++ LK + ES +
 ++++E +
 45 Sbjct: 670 DEKLLSIEDSKRDLES----
 LKEQLRAAQESKAKVEEGLKKLEEESSEKAELEKSKEM 725
 Query: 714 NRKLESVAEEHEILTKSYMELLQRN-ESTEKKNKDLQITCDSL-
 50 NKQIETVKKL NESLKE 771
 +KLES E +E KS ME ++++ E E+ K + +L +++ + +
 ++NES K+
 Sbjct: 726
 MKKLESTIESNETELKSSMETIRKSDEKLEQSKKSAEEDIKNLQHEKSDLISRINESEKD 785
 55 Query: 772 QNE-KSIAQLIEKEEQRKE-VQNLVDREHKL-
 ANLHQKTKVQEEKIKTLQKEREDKEET 828
 E KS ++ K E V+ +L + + K+ N + T V + K++
 ++++E +DK+

Sbjct: 786 IEELKSKLRIEAKSSSELETVKQELNNAQEKIRVNAEENT-
VLKSKLEDIERELKDKQAE 844

Query: 829 IDILR--KEL--SRTEQIRKEL-----SIKASSLEVQKAQLE-
5 GRLEEKESLVKLQ 874

I + KEL SR +++ +EL S + S EV+K Q+E
+L+EK L++ +

Sbjct: 845

IKSNQEEKELLTSRLKELEQELDSTQQAQKSEEEESRAEVRKFQVEKSQLEKAMLLET 904

Query: 875 QEEL-NK 880

+L NK

Sbjct: 905 YNDLVNK 911

15 Score = 173 (26.0 bits), Expect = 5.7e-09, P = 5.7e-09
Identities = 77/287 (26%), Positives = 146/287 (50%)

Query: 601 IEKLQSGMVVKDQICDVRISDINDVYEMKLSTLASKES--
20 RLQDLLETALALAQADRLI 658

++K + + K++ + IS + D E+ ST ES + D LE +
A+

Sbjct: 380 LKKYEEQIANKERQYNEEISQLND--EIT-
STQGENESIKKKNDELEGEVKAMKST---- 432

25 Query: 659 AQHRCQRTQAE TEARTLASMLREVERKNE--
ELSVLLKAQQVESERAQSDIEHLFQH-NR 715

++ + ++E +A L ++E+++KNE E S+L + +ESE + I+
L N

30 Sbjct: 433 SEEQSNLKKSEIDALNL--QIKELKKKNETNEASLLESIKSIESETVK--
IKELQDECNF 488

Query: 716 KLESVAEEHEILTKSY---

MELLQRNESTEKKNKDLQITCDSLNKQIETVKKLNESLKEQ 772

K + V+E + L S + L+ + +EK ++L L Q+E V

35 L+++ KE+

Sbjct: 489

KEKEVSELEDKLRKASEDKNSKYLELQKESEKIKEELDAKTTELKIQLEKVTNLSKA-KEK 547

40 Query: 773 NEKSIAQLIE-KEEQRKEVQNQL--VDREHKLAN--
LHQTQKVQEEKIKTLQKEREDKEE 827

+E +++L + E+RK + QL + E ++ N ++ K+ E T+
+E +K

Sbjct: 548

SESELSRLKKTSSSEERKNAEEQLEKLNKNEIQIKNQAFEKERKLLNEGSSTITQYSEKIN 607

45 Query: 828 TI-
DILRKELSRTEQIRKELSIKASSLEVQKAQLEGRLEEKESLVKLQQEELNKHSHMI 885

T+ D L + + E KE+ S LE + LEEK++ +K Q+E+
+ I

50 Sbjct: 608

TLEDELIRLQENELKAKEIDNTRSELEKVSLSNDELLEEKQNTIKSLQDEILSYKDKI 666

Score = 171 (25.7 bits), Expect = 9.3e-09, P = 9.3e-09
Identities = 76/311 (24%), Positives = 152/311 (48%)

55 Query: 596 NIEELIEKLQSGMVVKDQ-----
ICDVRISDINDVYEMKLSTLASKESRLQDLLET 648

N EE +EKL++ + +K+Q + + S I Y K++TL +

RLQ+ E KA

Sbjct: 565

NAEEQLEKLEKNEIQIKNQAFEKERKLLNEGSSTITQYSEKINTLEDELIRLQENELKA 624

5

Query: 649 LALAQADRLIAQHRCQRTQA-ETEARTLASMLREVERKNEELSVL-
LKAQQVESERAQSD 706

+ + + E + T+ S+ E+ +++++ K

+E + ++ D

10 Sbjct: 625

KEIDNTRSELEKVSLSNDELLEEKQNTIKSLQDEILSYKDKITRNDKLLSIED-SKRD 683

Query: 707 IEHLFQHNRL-ESVAEEHEILTKSYMELLQRNESTEKKN---
KDLQITCDS----LNKQ 758

15

+E L + R ES A+ E L K E + EK K L+ T +S

L

Sbjct: 684

LESLKEQLRAAQESKAKVEEGLKKLEEESSEKAELEKSKEMMKLESTIESNETELKSS 743

20 Query: 759 IETVKKLNESSLKEQNEKSIAQLIEK-

EEQRKEVQNLVDREHKLNLHOKTKVQEE---K 814

+ET++K +E L EQ++KS + I+ + ++ ++ +++ + E + L K

+++ + +

Sbjct: 744 METIRKSDEKL-

25 EQSKKSAEEDIKNLQHEKSDLISRINESEKDIEELKSKLRIEAKSSSE 802

Query: 815 IKTLQKEREDKEETIDILRKE----LSRTEQIRKELSIKASSL---
EVQKAQLEGRLEEK 867

++T+++E + +E I + +E S+ E I +EL K + + + +K

30 L RL+E

Sbjct: 803

LETVKQELNNAQEKIRVNAEENTVLKSKLEDIERELKDKQAEIKSNQEEKELLTSRLKEL 862

Query: 868 ESLVKLQEEELNK 880

35

E + Q++ K

Sbjct: 863 EQELDSTQKAQK 875

Score = 165 (24.8 bits), Expect = 4.1e-08, P = 4.1e-08

Identities = 65/286 (22%), Positives = 149/286 (52%)

40

Query: 595 LNIEELIEKLQSGMVVKDQICDVR-ISDIMDVYEMKLSTLASKESSL-
QDLLETALALA 652

+N ++ + L+ + K I +++ I++ ++ +++ + L+ ++ +

++L+E K+

45 Sbjct: 114 VNHQKETKSLKEDIAAK--

ITEIKAINENLEKMKIQCNLSKEKEHISKELVEYKS-RFQ 170

Query: 653 QADRLIAQHRCQRTQAETEARTLASMLREVERKNEELSVLLKAQQVESE---
-RAQSDIE 708

50

D L+A+ T+ + ++LA+ +++++ +NE L ++ + ES

Q+ I+

Sbjct: 171 SHDNLVAK----LTE---

KLKSLANNYKDMQAEENESLIKAVEESKNESSIQLSNLQNKID 223

55 Query: 709 HLFQH--NRKLE--

SVAEEHEILTKSYMELLQRNESTEKKNKDLQITCDSLQKQIETVKK 764

+ Q N ++E S+ + E L K+ +L Q E K+ + D

QI +K+

Sbjct: 224 SMSQEKENFQIERGSIIEKNIEQLKKTISDLEQTKEEIISKSDSSK---
DEYESQISLLKE 280

Query: 765

5 LNESLKEQNEKSIAQLIEKEEQRKEVQNLVDREHKLNLHQAQTKVQEEKIKTLQKERED 824
E+ N++++ ++ E + R+E++ +L ++ L K + E+ +K
+++ E

Sbjct: 281

KLETATTANDENVNKISELTKTREELEAELAAYKNLKNELETKLETSEKALKEVKENEH 340

10

Query: 825 -

KEETIDILRKELSRTEQIRKELSIKASSLEVQKAQLEGRLEEKESLVKLQAEELNK 880
KEE I L KE + T+Q L SLE + L +L++ E + ++
+ N+

15

Sbjct: 341 LKEEKIQ-

LEKEATETKQQLNSLRANLESLEKEHEDLAAQLKKYEEQIANKERQYNE 396

Score = 158 (23.7 bits), Expect = 1.9e-07, P = 1.9e-07

Identities = 74/268 (27%), Positives = 136/268 (50%)

20

Query: 598 EELIEKLQSGMVVKDQICDVRISDIMDVYEM--KL-

STLASKESRLQDLLET--KALALA 652
+E -K++ G+ ++ +++ EM KL ST+ S E+ L+ +ET
K+

25

Sbjct: 695

QESKAKVEEGLKKLEEESKEKAELEKSKEMMKLESTIESNETELKSSMETIRKSDEKL 754

Query: 653

QADRLIAQHRCQRTQAEARTLASMLREVERKNEELSVLLKAQQVESERAQSDIEHLFQ 712
+ + A+ + Q E L S + E E+ EEL L+ + S
++ + L

30

Sbjct: 755 EQSKKSAEEDIKNLQHEKS--

DLISRINESEKDIEELKSKLRIEAKSSSELETQVKQELNN 812

35

Query: 713 HNRKLESVAEEHEILTKSYMELLQRNESTEEKKNKDLQITCDSLKNQIET--
VKKLNESLK 770

K+ AEE+ +L KS +E ++R E K+K +I + K++ T
+K+L + L

40

Sbjct: 813 AQEKIRVNAEENTVL-KSKLEDIER----

ELKDKQAEIKSNQEEKELLTSRLKELEQELD 867

Query: 771 EQNEKSIAQLIEKEEQRKEVQNLVDREHKLNLHQAQTKVQEEKIKTLQKEREDKEE 827

+K AQ E EE R EV+ V++ + K L K K +

45

+++ + ++

Sbjct: 868 STQQK--AQKSE-

EESRAEVRKFQVEKSQLEKAMLLETQYNDLVNKEQAWKRDEDTVKK 924

Query: 828 TIDILRKELSRTEQIRKEL-SIKASSLEVQKAQLEGRLE 865

50

T D R+E+ E++ KEL ++KA + ++++A E R E

Sbjct: 925 TTDSQRQEI---EKLAKELDNLKAENSKLKEAN-EDRSE 959

Score = 155 (23.3 bits), Expect = 3.9e-07, P = 3.9e-07

Identities = 73/269 (27%), Positives = 133/269 (49%)

55

Query: 624 DVYEMKLSTLASKESRLQD-LLETQALALAQADRLIAQHRCQRTQAE--
EARTLASML 679

++ E K +T+ S LQD +L K ++L++ R + E+ +
 R
 Sbjct: 643 ELLEEKQNTIKS----
 LQDEILSYKDKITRNDKLLSIERDSKRDLKESLKEQLRAAQESK 698
 5 Query: 680 REVE---
 RKNEELSVLLKAQAVESERAQSDIEHLFQHNKLESVAEEHEILTKSYMELLQ 736
 +VE +K EE S KA+ +S+ +E + N + E +
 KS +L Q
 10 Sbjct: 699 AKVEEGLKKLEEESSEKAELEKSKEMMKLESTIESNET--
 ELKSSMETIRKSDEKLEQ 756
 Query: 737 RNESTEKKNKDLQITCDSLNKQIETVKKLNESLKEQ---
 NEKSIAQLIEKEEQRKEVQNQ 793
 15 ++S E+ K+LQ L +I +K E LK + KS ++L
 +++ Q +
 Sbjct: 757
 SKKSAEEDIKNLQHEKSDLISRINESEKDIEELKSKLRIEAKSSSELETVKQELNNAQEK 816
 20 Query: 794 L-VQREH-----
 KLANLHQKTKVQEEKIKTLQKEREDKEETIDILRKELSRTEQIRKEL 846
 + V+ E KL ++ ++ K ++ +IK+ Q+E+E + L +EL
 T+Q + +
 Sbjct: 817
 25 IRVNAEENTVLKSKLEDIERELKDKQAEIKSNQEEKELLTSRLKELEQELDSTQQ-KAQK 875
 Query: 847 SIKASSLEVQKAQLE-GRLEEKESLVKLQEEEL-NK 880
 S + S EV+K Q+E +L+EK L++ + +L NK
 Sbjct: 876 SEESRAEVRKFQVEKSQLEKAMLLETKYNDLVNK 911
 30 Score = 146 (21.9 bits), Expect = 3.5e-06, P = 3.5e-06
 Identities = 73/311 (23%), Positives = 152/311 (48%)
 Query: 520 DNREQVQSGLRIL-----LEAAPLPDFPALV--
 35 LGESIAANNAYRQQETEHIPRK-MPWQ 571
 +++ +V+ GL+ L E A L ++ L +I +N + E I
 + +
 Sbjct: 696
 ESKAKVEEGLKKLEEESSEKAELEKSKEMMKLESTIESNETELKSSMETIRKSDEKLE 755
 40 Query: 572 SSNHSFPTS IKCLTPHLKDGVPGLNIEEL-
 IEKLQSGMVVKDQICDVRISDINDVYEMKL 630
 S S IK L D + +N E IE+L+S + + + + S
 ++ + +L
 45 Sbjct: 756 QSKKSAEEDIKNLQHEKSDLISRINESEKDIEELKSKLRI-----
 EAKSSSELETVKQEL 810
 Query: 631 STLASK---
 50 ESRLQDLLETALALAQADRLIAQHRCQRTQAETEARTLASMLREVERKNE 687
 + K + +L++K L +R + + + + E L S
 L+E+E++ +
 Sbjct: 811 NNAQEKIRVNAEENTVLKSK---
 LEDIERELKDKQAEIKSNQEEKELLTSRLKELEQELD 867
 55 Query: 688
 ELSVLLKAQAVESERAQSDIEHLFQHNKLESVAEEHEILTKSYMELLQARNESTEKKND 747
 S KAQ+ E E +++++ FQ + + E+ +L Y +L+ +
 ++ ++

Sbjct: 868 --STQQAQKSEEE-SRAEVRK-FQVEKS--
QLDEKAMLLKETKYNDLVNKEQAWKRDED 921

5 Query: 748 LQITCDSL NKQIETVKKL NESLKEQNEKSIAQLIEKEEQRKEVQNLV---
DREHKL ANL 804
++ T DS ++IE + K ++LK +N K L E E R E+ + ++ D
+ K N

10 Sbjct: 922 VKKTTDSQRQEI EK LAKEL DNLKAENSK----
LKEANEDRSEIDDLMLLVTDLDEK--NA 975

Query: 805 HQKTKVQEEKIKTLQKEREDKEETID 830
++K+++ ++ E +D+EE D
Sbjct: 976 KYRSKLKDLGVEISSDEEDDEEEEDD 1001

15 Score = 146 (21.9 bits), Expect = 4.6e-06, P = 4.6e-06
Identities = 82/313 (26%), Positives = 145/313 (46%)

20 Query: 598
EELIEKLQSGMVVKDQICDVRISDIMDVYEMKLSTLASKESRLQDLLET KALALAQAADRL 657
EEL +L + +K+++ + + E+K + KE ++Q LE +A
Q
Sbjct: 304 EELEAELAAYKNL KNELET KLETSEKALKEVKENEHLKEEKIQ--
LEKEATETKQ--- 358

25 Query: 658 IAQHRCQRTQAETEARTLASMLREVERK-----NEELSVL---
LKAQQVESERAQSD 706
+ R E E LA+ L++ E + NEE+S L + + Q
E+E +

30 Sbjct: 359
LNSLRANLESLEKEHEDLAAQLKKYEEQIANKERQYNEEISQLNDEITSTQQENESIKKK 418

Query: 707 IEHLFQHNRLKLESVAEEHEILTKSYMELLQRN-
ESTEKKNKDLQITCDSL NKQIET-VKK 764
+ L + ++S +EE L KS ++ L + +KKN+ + + K

35 IE+ K
Sbjct: 419
NDELEGEVKAMKSTSEEQSNLKKSEIDALNLQIKELKKKNETNEASLLESIKSIESETVK 478

40 Query: 765 LNESLKEQN--EK SIAQLIEK---EEQRKEVQNLVDREHKL AN-LHQKT--
-KVQEEKI 815
+ E E N EK +++L +K E + +L K+ L KT
K+Q EK+

45 Sbjct: 479
IKELQDECNFKEKEVSELEDK LKASEDKNSKYLELQKESEKIKEELDAKTTELKIQLEKV 538

Query: 816
KTLQKEREDKEETIDILRKELSRTEQIRKELSIKASSLEVQKAQLEGRLEEKESLVKLQ 875
L K +E E ELSR ++K S + + E Q +L+ ++ K
+ ++

50 Sbjct: 539 TNL SKAKEKSES-----ELSR---
LKKTSSEERKNAEEQLEKLKNEIQIKNQAFER 588

Query: 876 EELNKHSHMIAMIHSLSGGKINPETVNL 903
+ LN+ S I +S + E + L
55 Sbjct: 589 KLLNEGSSITQ EYSEKINTLEDELIRL 616

Score = 145 (21.8 bits), Expect = 5.9e-06, P = 5.9e-06
Identities = 59/246 (23%), Positives = 115/246 (46%)

Query: 634 ASKESRLQ-

DLLETKALALAQAADRLIAQHRCQRTQAETEARTLASMLREVERKNEELSVL 692

+ ES +Q L+ K +++Q + +R E L + ++E+

5 EE ++

Sbjct: 207 SKNESSIQLSNLQNKIDSMSQKEKE---

NFQIERGSIEKNIEQLKKTISDLEQTKEE--II 261

Query: 693 LKAQQVESERAQSDIEHLFQHNRLKLESVAEEHEI-----

10 LTKSYMELLQRNESTTEKKNKD 747

K+ + E +S I L + + + A + + LTK+ EL

+ + +

Sbjct: 262 SKSDSSKDEY-ESQIS-

LLKEKLETATTANDENVNKISELTKTREELEAELAAYKNLKNE 319

15

Query: 748

LQITCDSLNKQIETVKKL NESLKEQNEKSIAQLIEKEEQRKEVQNLVDREHKLANLHQK 807

L+ ++ K ++ VK+ E LKE+ + + E ++Q ++ L E

+ +L +

20 Sbjct: 320

LETKLETSEKALKEVKENEEHLKEEKIQLEKEATETKQQLNSLRANLESLEKEHEDLAAQ 379

Query: 808

TKVQEEKIKTLQKEREDKEETIDILRKELSRTEQIRKELSIKASSLEVQKAQLEGRLEEK 867

K EE+I KER+ EE I L E++ T+Q + + K LE +

25

++ EE+

Sbjct: 380 LKKYEEQIAN--KERQYNEE-

ISQLNDEITSTQENESIKKKNDELEGEVKAMKSTSEEQ 436

30 Query: 868 ESLVKLQEEELN 879

+L K + + LN

Sbjct: 437 SNLKKSEIDALN 448

Score = 137 (20.6 bits), Expect = 4.2e-05, P = 4.2e-05

35 Identities = 81/312 (25%), Positives = 140/312 (44%)

Query: 598 EELIEKLQSGMVVKDQICDVRISDIMDVYEMKLSTLASK-ESRLQDLLET-
KALALAQAQ 655

+EL ++++ ++ +++ S+I D +++ L K E+ LLE+

40 K++

Sbjct: 420 DELEGEVKAMKSTSEEQSNLKKSEI-

DALNLQIKELKKKNETNEASLLESIKSIESETVK 478

Query: 656

45 RLIAQHRCQRTQAETEARTLASMLREVERKNEELSVLLKAQQVESERAQSDIEHLFQHNRL 715

Q C E E L L+ E KN + L K + E +

L

Sbjct: 479 IKELQDECNFK--

EKEVSELEDKLGASEDKNSKYLELQKESEKIKEELDAKTTELKIQLE 536

50

Query: 716

KLESVAEEHEILTKSYMELLQRNESTTEKKNKDLQITCDSLNKQIETVKKL NESLKEQNEK 775

K+ ++++ E ++S + L++ S E+KN + Q+ QI+ + +

K NE

55 Sbjct: 537 KVTNLSKAKE-KSESELSRLKKTSSSEERKNAEEQLEKLKNEIQIKN-

QAFEKERKLLNEG 594

Query: 776 SIAQLIEKEEQRKEVQNLV--DREHKL-ANLHQKTKVQEEKIKTLQKER-
EDKEETIDI 831

S E E+ ++++L+ E++L A T+ + EK+ E
E+K+ TI

5 Sbjct: 595
SSTITQYSEKINTLEDELIRLQENELKAKEIDNTRSELEKVSLSNDELLEEKQNTIKS 654

Query: 832 LRKE-LSRTEQI----RKELSIKASS---LEVQKAQLEGRLEEK----
ESLVKLQQE--- 876

10 L+ E LS ++I K LSI+ S LE K Q L E K E L
KL++E

Sbjct: 655
LQDEILSYKDKITRNDKLLSIEDSKRDLESLEQLRAAQESKAKVEEGLKKLEEESK 714

15 Query: 877 ---ELNKHSHMIAMIHS 890

EL K M+ + S

Sbjct: 715 EKAELEKSKEMMKLES 731

Score = 128 (19.2 bits), Expect = 3.9e-04, P = 3.9e-04

20 Identities = 80/356 (22%), Positives = 148/356 (41%)

Query: 546 LGESIAANNAYRQQETEHIPRKMPWQSSNHSFPTSIKCLTPHL-----
KDGVPGLN-I 597

L E + ++ E+ + ++S+ H SIK L L K

25 G+N +

Sbjct: 25
LDEMTQLRDVLETQDKENQATALLEYKSTIHKQEDSIKTLEKELETILSQKKAEDGINKM 84

Query: 598 EELIEKLQSGMVVKDQICD--

30 VRISDINDVYEMKLSTLASKESRLQDLLETKALALAQAD 655

+ + L M ++ C + D +V K T + KE + E

KA+ +

Sbjct: 85 GKDLFALSREMQAVEENCKNLQKEKDKSNVNHQK-
ETKSLKEDIAAKITEIKAIN-ENLE 142

35

Query: 656

RLIAQHRCQRTQAEARTLASMLREVERKNEELSULLKAQQVESERAQSDIEHLFQHNH 715
++ Q C E E ++ L E + + + L+ + + ++

+ + N

40 Sbjct: 143 KMKIQ--CNNLSKEKEH--

ISKELVEYKSRFQSHDNLVAKLTEKLKSLANNYKDMQAEENE 198

Query: 716 KLESVAEEHEILTKSYMELLQRN-
ESTEKKNKDLQITCDLNLKQIETVKKLNESEKQNE 774

45 L EE + + + LQ +S ++ ++ QI S+ K IE +KK

L++ E

Sbjct: 199

SLIKAVEESKNESSIQLSNLQNKIDSMSQEKENFQIERGSIEKNIEQLKKTISDLEQTK 258

50 Query: 775

KSIAQLIEKEEQRKEVQNLVDREHKLANLHQKTKVQEEKIKTLQKEREDKEETI----- 829
+ I++ + + E ++Q+ + KL KI L K RE+ E

+

Sbjct: 259 EIISK---

55 SDSSKDEYESQISLLKEKLETATTANDENVNKISELTKTREELEAELAAYKN 315

Query: 830 --DILRKELSRTEQIRKELSIKASSLEVQKAQLEGRLEE-KESLVKLQ--
EELNK-HSH 883

+ L +L +E+ KE+ L+ +K QLE E K+ L L+ E

L K H

Sbjct: 316

LKNELETKLETSEKALKEVKENEEHLKEEKIQLEKEATETKQQLNSLRANLESLEKEHED 375

5

Query: 884 MIAMI 888

+ A +

Sbjct: 376 LAAQL 380

10 Score = 117 (17.6 bits), Expect = 3.8e-03, P = 3.8e-03
Identities = 50/240 (20%), Positives = 111/240 (46%)

Query: 634

ASKESRLQDLLETALALAQADRLIAQHRCQRTQAETEARTLASMLREVERKNEELSVLL 693

15

A E L+ L E + A+ ++ + + E+ L S + + +

+E+L

Sbjct: 699

AKVEEGLKKLEEESSEKAELEKSKEMMKLESTIESNETELKSSMETIRKSDEKLEQSK 758

20 Query: 694 KAQQVESERAQ---SD-

IEHLFQHNRLKLESVAEEHEILTKSYMELLQNESTEEKNKDLQ 749

K+ + + + Q SD I + + + +E + + I KS EL +

+ ++

Sbjct: 759

25 KSAEEDIKNLQHEKSDLISRINESEKDIEELKSKLRIEAKSSSELETVKQELNNAQEKIR 818

Query: 750

ITCDSLNKQIETVKKLNESEKQNEKSIAQLIEKEEQRKEVQNLVDREHKLNLHQQTK 809

+ + N +++ KL + +E +K A++ +E+++ + ++L + E +L

30

+ QK +

Sbjct: 819 VNAEE-NTVLKS--KLEDIERELKDKQ-

AEIKSNQEEKELLTSRLKELEQELDSTQQAQ 874

Query: 810 VQEEK----

35 IKTLQKEREDKEETIDILRKELSRTEQIRKELSIKASSLEVQKAQLEGRLE 865

EE+ ++ Q E+ +E +L E + + KE + K V+K

+ + +

Sbjct: 875 KSEESRAEVRKFQVEKSQLEKAMLL--

ETKYNDLVNKEQAWKRDEDTVKKTT-DSQRQ 931

40

Query: 866 EKESLVK 872

E E L K

Sbjct: 932 EIEKLAK 938

45 Score = 109 (16.4 bits), Expect = 2.6e-02, P = 2.5e-02
Identities = 64/284 (22%), Positives = 135/284 (47%)

Query: 598

EELIEKLQSGMVVKDQICDVRI SDIMDVYEMKLSTLASKESRLQDLLETALALA---QA 654

50

+E+++KL+S + + + I E + S E +++L K+

++ ++

Sbjct: 723

KEMMKLESTIESNETELKSSMETIRKSDEKLEQSKKSAEEDIKNLQHEKSDLISRINES 782

55 Query: 655 DRLIAQHRCQ-

RTQAETEARTLASMLREVERKNEELSVLLKAQQVESERAQSDIEH-LFQ 712

++ I + + + R +A++ + L ++ +E+ E++ V + V + +

DIE L

Sbjct: 783 EKDI EELKSKLR IEAKSSSE-LET VKQELNNAQ EKIRVNAEENTVLKSKLE-
DIERELKD 840

Query: 713 HNRKLESVAEEHEILTKSYMELLQ RNESTEKK-NKDLQITCDSL NK-
5 QIETVKKL NES-- 768

+++S EE E+LT EL Q +ST++K K + + + K Q+E
+L+E

Sbjct: 841 KQAEIKSNQEEKELLTSRLKELEQELDSTQQAQKSEESRAEVRKFQVEK-
S QLDEKAM 899

10

Query: 769 LKEQNEKSIA---Q LIEKEEQ--
RKEVQNLVDREHKL ANLH QKTKVQEEKIKTLQKERE 823

L E + Q +++E +K +Q + E KLA K +
K+K ++R

15

Sbjct: 900 LLET KYNDLVNKEQAWKRDEDTVKKTTDSQRQ EIE-
KLAKELDNLKAENSKLKEANEDRS 958

Query: 824 DKEETI----DILRKELSRTEQIRKELSIKASSLEVQKAQLEGRLEEKE
868

20

+ ++ + D+ K ++ K+L ++ SS E + E E+ E
Sbjct: 959 EIDDLMLLVTDLDEKNAKYRSKL-KDLGVEISSDEEDDEEEEDDEEDDE
1006

Score = 96 (14.4 bits), Expect = 1.1e+00, P = 6.6e-01
25 Identities = 40/210 (19%), Positives = 101/210 (48%)

Query: 681 EVERKN--
EELS VLLKAQQVESERAQSDIEHLFQ HNRKLESVAEEHEILTKSYMELLQ RN 738

30

E E KN + L + + + V + + + L ++ + + + L K

+L +

Sbjct: 15

ETELKNVRDSLDEMTQLRDVLETKDKENQ TALLEYKSTIHKQEDSIKTLEKELETILSQK 74

Query: 739 ESTE----

35

KKNKDLQITCDSL NKQIETVKKL NESLKEQNEKSIAQ LIEKEEQ RKEVQNL 794

+ E K KDL +L++++ V++ ++L+++ +KS + +++

K ++ +

Sbjct: 75 KKAEDGINKMGKDLF-----ALSREMQAVEENCKNLQKEKDKSN---

VNHQKETKSLKEDI 127

40

Query: 795 VDREHKL ANLH QKTKVQEEKIKTLQKERE D-
KEETIDILRKELSRTEQIRKELSIKASSL 853

+ ++ +++ + + + L KE+E +E ++ + S + K

L+ K SL

45

Sbjct: 128 AAKITEIKAINENLEKMKIQCNNLSKEKEHISKELVEYKSRFQSHDNLVAK-
LTEKLKSL 186

Query: 854 EVQKAQLEGRLEEKESLVKLQ QEELNKHSHMIAMIHS 890

++ E ESL+K +E N+ S ++ + +

50

Sbjct: 187 ANNYKDMQA---ENESLIKAVEESKNESSIQLSNLQ N 220

Score = 52 (7.8 bits), Expect = 2.0e-10, P = 2.0e-10
Identities = 39/167 (23%), Positives = 74/167 (44%)

55

Query: 99 LNSVLAGVVCRSSHTDSVFLQ CIQLLQKLTYNVKIFYSGANIDEL-
ITFLIDHIQSS EDE 157

LN + + ++ ++ L+ I+ ++ T +K N E ++ L D
+++SED+

Sbjct: 447
LNLQIKELKKKNETNEASLLESIKSIESETVKIKELQDECNFKEKEVSELEDKCLKASEDK 506

Query: 158 -
5 LKMPCLGLLANLCRHNLSVQTHIKTLSNVKSFYRTLITLLAHSSLTVVVFALSILSSLT 216
K L + + L +T T ++ T ++ S + +
S

Sbjct: 507 NSKYLELQKESEKIKEELDAKT---
10 TELKIQLKVTNLSKAKEKSESELSRLKKTSSSEER 563

Query: 217 LN-EEVGKELFHARNI-HQTFQLIFNILINGDGTLTRKYS--VDLLMDLL
262

Sbjct: 564 KNAEEQLKLEKLNKNEIQIKNQAFEKERKLLNEGSSTITQYSEKINTLEDEL
15 613
N EE EKL + I +Q F+ +L G T+T++YS ++ L D L

Pedant information for DKFZphm12_12j1, frame 2

Report for DKFZphm12_12j1.2

25 [LENGTH] 905
[MW] 102067.81
[pI] 5.85
[HOMOL] TREMBL:SCINTANA_1 Saccharomyces cerevisiae
integrin analogue gene, complete cds. 1e-14
30 [FUNCAT] 08.07 vesicular transport (golgi network, etc.) [S.
cerevisiae, YDL058w] 5e-16
[FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae,
YDL058w] 5e-16
[FUNCAT] 1 genome replication, transcription, recombination and
repair [M. jannaschii, MJ1322] 1e-10
35 [FUNCAT] 09.10 nuclear biogenesis [S. cerevisiae, YDR356w]
2e-10
[FUNCAT] 30.04 organization of cytoskeleton [S. cerevisiae,
YDR356w] 2e-10
[FUNCAT] 03.22 cell cycle control and mitosis [S. cerevisiae,
40 YDR356w] 2e-10
[FUNCAT] 30.10 nuclear organization [S. cerevisiae, YKR095w]
1e-09
[FUNCAT] 11.04 dna repair (direct repair, base excision repair
and nucleotide excision repair) [S. cerevisiae, YKR095w] 1e-09
45 [FUNCAT] 08.22 cytoskeleton-dependent transport [S. cerevisiae,
YHR023w MY01 - myosin-1 isoform] 4e-09
[FUNCAT] 03.04 budding, cell polarity and filament formation
[S. cerevisiae, YHR023w MY01 - myosin-1 isoform] 4e-09
[FUNCAT] 03.25 cytokinesis [S. cerevisiae, YHR023w MY01 -
50 myosin-1 isoform] 4e-09
[FUNCAT] 99 unclassified proteins [S. cerevisiae, YNL091w]
3e-08
[FUNCAT] 09.25 vacuolar and lysosomal biogenesis [S.
cerevisiae, YOR326w] 1e-08
55 [FUNCAT] 08.16 extracellular transport [S. cerevisiae,
YOR326w] 1e-08
[FUNCAT] 09.13 biogenesis of chromosome structure [S.
cerevisiae, YLR086w] 8e-08

- [FUNCAT] 98 classification not yet clear-cut [S. cerevisiae, YJR134c] 1e-07
 [FUNCAT] 06.07 protein modification (glycosylation, acylation, myristylation, palmitoylation, farnesylation and processing) [S. cerevisiae, YKL201c] 4e-07
 5 [FUNCAT] 30.05 organization of centrosome [S. cerevisiae, YIL144w] 4e-06
 [FUNCAT] 03.07 pheromone response, mating-type determination, sex-specific proteins [S. cerevisiae, YNL079c] 5e-06
 10 [FUNCAT] 03.01 cell growth [S. cerevisiae, YNL079c] 5e-06
 [FUNCAT] 08.99 other intracellular-transport activities [S. cerevisiae, YNL079c] 5e-06
 [FUNCAT] 09.04 biogenesis of cytoskeleton [S. cerevisiae, YKL179c] 6e-06
 15 [FUNCAT] 30.02 organization of plasma membrane [S. cerevisiae, YER008c] 8e-06
 [FUNCAT] 03.19 recombination and dna repair [S. cerevisiae, YNL250w] 1e-05
 [FUNCAT] 03.13 meiosis [S. cerevisiae, YDR285w] 1e-05
 20 [FUNCAT] 30.13 organization of chromosome structure [S. cerevisiae, YDR285w] 1e-05
 [FUNCAT] 11.01 stress response [S. cerevisiae, YPR141c] 2e-05
 [FUNCAT] 06.10 assembly of protein complexes [S. cerevisiae, YPR141c] 2e-05
 25 [FUNCAT] 06.01 protein folding and stabilization [S. cerevisiae, YNL227c] 9e-05
 [FUNCAT] 05.04 translation (initiation, elongation and termination) [S. cerevisiae, YAL035w] 1e-04
 [FUNCAT] 10.05.99 other pheromone response activities [S. cerevisiae, YHR158c] 1e-04
 30 [FUNCAT] 0 chaperones [M. genitalium, MG355] 2e-04
 [FUNCAT] 03.22.01 cell cycle check point proteins [S. cerevisiae, YGL086w] 2e-04
 [FUNCAT] 03.10 sporulation and germination [S. cerevisiae, YNL225c] 3e-04
 35 [FUNCAT] r general function prediction [M. jannaschii, MJ1254] 4e-04
 [FUNCAT] 08.01 nuclear transport [S. cerevisiae, YPL174c] 4e-04
 [FUNCAT] 04.05.01.01 general transcription activities [S. cerevisiae, YMR227c TAF67 - TFIID subunit] 6e-04
 40 [BLOCKS] PRO1002E
 [BLOCKS] BL011608 Kinesin light chain repeat proteins
 [BLOCKS] BL00326 Tropomyosins proteins
 [SCOP] d2tmab_1.105.4.1.1 Tropomyosin [rabbit
 45 (Oryctolagus cuniculus) 3e-23
 [EC] 3.6.1.32 Myosin ATPase 4e-10
 [PIRKW] nucleus 5e-09
 [PIRKW] phosphotransferase 2e-07
 [PIRKW] blocked amino end 1e-06
 50 [PIRKW] duplication 2e-07
 [PIRKW] citrulline 3e-08
 [PIRKW] tandem repeat 4e-10
 [PIRKW] heterodimer 1e-07
 [PIRKW] heart 4e-08
 55 [PIRKW] endocytosis 7e-08
 [PIRKW] transmembrane protein 1e-14
 [PIRKW] serine/threonine-specific protein kinase 2e-07
 [PIRKW] cell wall 2e-06

	[PIRKW]	zinc finger 7e-08
	[PIRKW]	DNA binding 3e-09
	[PIRKW]	metal binding 7e-08
	[PIRKW]	muscle contraction 4e-10
5	[PIRKW]	brain 2e-06
	[PIRKW]	acetylated amino end 2e-07
	[PIRKW]	heterotetramer 5e-07
	[PIRKW]	actin binding 4e-10
	[PIRKW]	mitosis 1e-08
10	[PIRKW]	microtubule binding 1e-08
	[PIRKW]	ATP 4e-10
	[PIRKW]	chromosomal protein 1e-07
	[PIRKW]	thick filament 9e-10
	[PIRKW]	phosphoprotein 1e-09
15	[PIRKW]	skeletal muscle 1e-08
	[PIRKW]	calcium binding 3e-08
	[PIRKW]	alternative splicing 9e-10
	[PIRKW]	DNA condensation 1e-07
	[PIRKW]	coiled coil 1e-14
20	[PIRKW]	P-loop 2e-10
	[PIRKW]	heptad repeat 5e-09
	[PIRKW]	methylated amino acid 4e-10
	[PIRKW]	immunoglobulin receptor 2e-07
	[PIRKW]	peripheral membrane protein 7e-08
25	[PIRKW]	cardiac muscle 4e-08
	[PIRKW]	hydrolase 4e-10
	[PIRKW]	microtubule 5e-09
	[PIRKW]	muscle 4e-08
	[PIRKW]	membrane protein 5e-09
30	[PIRKW]	EF hand 3e-08
	[PIRKW]	cell division 1e-06
	[PIRKW]	cytoskeleton 6e-09
	[PIRKW]	hair 3e-08
	[PIRKW]	calmodulin binding 7e-08
35	[PIRKW]	Golgi apparatus 2e-07
	[SUPFAM]	hypothetical protein YJL074c 5e-09
	[SUPFAM]	unassigned Ser/Thr or Tyr-specific protein kinases 2e-07
	[SUPFAM]	myosin motor domain homology 2e-10
40	[SUPFAM]	alpha-actinin actin-binding domain homology 6e-09
	[SUPFAM]	tropomyosin 2e-08
	[SUPFAM]	kinesin heavy chain 5e-07
	[SUPFAM]	plectin 6e-09
	[SUPFAM]	SAM homology 1e-06
45	[SUPFAM]	trichohyalin 3e-08
	[SUPFAM]	ribosomal protein S10 homology 6e-09
	[SUPFAM]	protein kinase C zinc-binding repeat homology 5e-09
	[SUPFAM]	giantin 7e-08
	[SUPFAM]	protein kinase homology 2e-07
50	[SUPFAM]	protein 4.1 membrane-binding domain homology 9e-08
	[SUPFAM]	human early endosome antigen 1 7e-08
	[SUPFAM]	myosin MY02 2e-06
	[SUPFAM]	M5 protein 3e-09
	[SUPFAM]	Mycoplasma genitalium hypothetical protein MG218 5e-09
55	[SUPFAM]	myosin heavy chain 2e-10
	[SUPFAM]	conserved hypothetical P115 protein 3e-09
	[SUPFAM]	centromere protein E 1e-08
	[SUPFAM]	calmodulin repeat homology 3e-08

10

15

20

25

30

35

40

45

50

55

MEM

5 SEQ EVLLTLCGDDTLKMHIAKILTTVKCTTLIEQQFTYGKIDLGFGTKVADSELCKLAADVIL
SEG
PRD hhhhhhccccchhhhhhhhhhhheeeeeeeeeecccccccccceehhhhhhhhhhhhh
COILS

10 MEM

10 SEQ KTLDLINKLKPLVPGMEVSFYKILQDPRLITPLAFALTSDNREQVQSGLRILLEAAPLPD
SEG
PRD hhhhhhccccccccccccceecccccchhhhhhccccchhhhhhhhhhhhhcccc
COILS

15 MEM

20 SEQ FPALVLGESIAANNAYRQQETEHIPRKMPWQSSNHSFPTSIKCLTPHLKDGVPGLNIEEL
SEG
PRD ceeeeehhhhhhhhhhhhhhhhhhhhhhhhhccccccccchhhhhhhhhhhhhhhhhhhhh
COILS

25 MEM

25 SEQ IEKLQSGMVVKDQICDVRISDIMDVYEMKLSTLASKESRLQDLLETALALAQADRLIAQ
SEG
PRD hhh
COILS

30 MEM

30 SEQ HRCQRTQAETEARTLASMLREVERKNEELSVLLKAQQVESERAQSDIEHLFQHNKLESV
SEG
PRD hhh
COILS

35 MEMCC.....

40 SEQ AEEHEILTKSYMELLQARNESTEKKNKDLQITCDSLNKQIETVKKLNESLKEQNEKSIAQL
SEG
PRD hhh
COILSCC

45 MEM

45 SEQ IEKEEQRKEVQNLVDREHKLNLHAKTKVQEEKIKTLQKEREDKEETIDILRKELSRTE
SEG
PRD hhh
COILSCC

50 MEM

55 SEQ QIRKELSIKASSLEVQKAQLEGRLEEKESLVKLQEEELNKHSHMIAMIHSLSGGKINPET
SEG
PRD hhh
COILSCC

MEM

SEQ VNLSI
SEG
PRD ccccc
COILS
MEM

5

Prosites for DKFZphm12_12j1.2

10

PS00029

331->353

LEUCINE_ZIPPER

PD0C00029

(No Pfam data available for DKFZphm12_12j1.2)

15

DKFZphmel2_7g14

5 group: intracellular transport and trafficking

DKFZphmel2_7g14 encodes a novel 973 amino acid protein with similarity to the dor (deep orange) protein of drosophila melanogaster.

10

The novel protein is also similar to the vakuolar membrane protein pep3 of *Saccharomyces cerevisiae*, which is involved in protein sorting mechanisms. The expression profile is ubiquitous and a role in protein transport/targeting is likely.

15

The new protein can find application in modulation of the sorting of proteins into different compartments.

20

similarity to DEEP ORANGE (*Drosophila melanogaster*)

perhaps complete cds. and full length

Sequenced by MediGenomix

25

Locus: unknown

Insert length: 3951 bp

Poly A stretch at pos. 3893, polyadenylation signal at pos. 3874

30

```
1  GCGGCGCTCA  CGGGGGCGGG  AGTCAGCTGA  GCTGCCGGGG  CGAGGTTGGG
51  ATCACCTGGC  ACCGGCTGAA  GGGAGCCTGT  GATTTTTTTT  TAGCGGGGGC
101 GGGGAGTAAG  GTGCAAGACT  GCGCCAGATT  CAAGGACGAG  GGCTGCCCCG
35 151 TTATCTCGCT  GCATAAGGCA  AGAGCAAGAG  GATCCTCAGG  ATTTTAAAGA
201 GGAGGCGACG  GCTGCAGGTT  CCCAGGATCT  GTCAGAGGCT  GGGGAGTTAC
251 AGCTTCCATT  CTGGGGGCGA  GGGGACCCCG  GGGGGGTAGC  CCTTTGTAA
301 TCCCCAGGCC  CCGGACAAAG  AGCCCAGAGG  CCGGGCACCA  TGGCGTCCAT
351 CCTGGATGAG  TACGAGAACT  CGCTGTCCCG  CTCGGCCGTC  TTGCAGCCCG
40 401 GCTGCCCTAG  CGTGGGCATC  CCCCAGTCTG  GGTATGTGAA  TGCCAGCTG
451 GAGAAGGAAG  TGCCCATCTT  CACAAAGCAG  CGCATTGACT  TCACCCCTTC
501 CGAGCGCATT  ACCAGTCTTG  TCGTCTCCAG  CAATCAGCTG  TGCATGAGCC
551 TGGGCAAGGA  TACACTGCTC  CGCATTGACT  TGGGCAAGGC  AAATGAGCCC
601 AACCACGTGG  AGCTGGGACG  TAAGGATGAC  GCAAAAGTTC  ACAAGATGTT
45 651 CCTTGACCAT  ACTGGCTCTC  ACCTGCTGAT  TGCCCTGAGC  AGCACGGAGG
701 TCCTCTACGT  GAACCGAAAT  GGACAGAAGG  TACGGCCACT  AGCACGCTGG
751 AAGGGGCGAG  TGGTGGAGAG  TGTGGGTTGG  AACAAGGCAC  TGGGCACGGA
801 GAGCAGCACA  GGCCCCATCC  TGGTCGGGAC  TGCCCCAAGG  CACATCTTTG
851 AAGCAGAGCT  CTCAGCCAGC  GAAGGTGGGC  TTTTCGGCCC  TGCTCCGGAT
50 901 CTCTACTTCC  GCCCATTGTA  CGTGCTAAAT  GAAGAAGGGG  GTCCAGCACC
951 TGTGTGCTCC  CTTGAGGCCG  AGCGGGGCCC  TGATGGGCGT  AGCTTTGTTA
1001 TTGCCACCAC  TCGGCAGCGC  CTCTTCCAGT  TCATAGGCCG  AGCAGCAGAG
1051 GGGGCTGAGG  CCCAGGGTTT  CTCAGGGCTC  TTTGCAGCTT  ACACGGACCA
1101 CCCACCCCCA  TTCCGTGAGT  TTCCCAGCAA  CCTGGGCTAC  AGTGAGTTGG
55 1151 CCTTCTACAC  CCCCAGCTG  CGCTCCGCAC  CCCGGGCCTT  CGCCTGGATG
1201 ATGGGGGATG  GTGTGTTGTA  TGGGGCATTG  GACTGTGGGC  GCGCTGACTC
1251 TCTGCTGAGC  GAGGAGCGAG  TCTGGGAGTA  CCCAGAGGGG  GTAGGGCCTG
1301 GGGCCAGCCC  ACCCCTAGCC  ATCGTCTTGA  CCCAGTTCCA  CTTCCTGCTG
```

1351 CTACTGGCAG ACCGGGTGGA GGCAGTGTGC AACTGACCG GGCAGGTGGT
1401 GCTGCGGGAT CACTTCCTGG AGAAATTTGG GCCGCTGAAG CACATGGTGA
1451 AGGACTCCTC CACAGGCCAG CTGTGGGCCT AACTGAGCG GGCTGTCTTC
1501 CGCTACCACG TGCAACGGGA GGCCCGAGAT GTCTGGCGCA CCTATCTGGA
5 1551 CATGAACCGC TTCGATCTGG CCAAAGAGTA TTGTGAGAG CGGCCCGACT
1601 GCCTGGACAC GGTCTTGCC CGGGAGGCCG ATTTCTGCTT TCGCCAGCGT
1651 CGCTACCTGG AGAGCGCACG CTGCTATGCC CTGACCCAGA GCTACTTTGA
1701 GGAGATTGCC CTCAAGTTCC TGGAGGCCCG ACAGGAGGAG GCTCTGGCTG
1751 AGTTCCTGCA GCGAAACTG GCCAGTTTGA AGCCAGCCGA ACGTACCCAG
10 1801 GCCACACTGC TGACCACCTG GCTGACAGAG CTCTACCTGA GCCGGCTTGG
1851 GGCTCTGCAG GGCAGCCAG AGGCCCTGAC TCTCTACCGA GAAACCAAGG
1901 AATGCTTTTCG AACCTTCCCTC AGCAGCCCCC GCCACAAAGA GTGGCTCTTT
1951 GCCAGCCGGG CCTCTATCCA TGAGCTGCTC GCCAGTCATG GGGACACAGA
2001 ACACATGGTG TACTTTGCAG TGATCATGCA GGACTATGAG CGGGTGGTGG
15 2051 CTTACCACTG TCAGCACGAG GCCTACGAGG AGGCCCTGGC CGTGCTCGCC
2101 CGCCACCGTG ACCCCAGCT CTTCTACAAG TTCTCACCCA TCCTCATCCG
2151 TCACATCCCC CGCCAGCTTG TAGATGCCTG GATTGAGATG GGCAGCCGGC
2201 TGGATGCTCG TCAGCTCATT CCTGCCCTGG TGAACACAG CCAGGGTGGT
2251 GAGGTCCAGC AGGTGAGCCA GGCCATCCGC TACATGGAGT TCTGCGTGAA
20 2301 CGTGCTGGGG GAGACTGAGC AGGCCATCCA CAACTACCTG CTGTCACTGT
2351 ATGCCCCGTGG CCGGCCGGAC TCACTACTGG CCTATCTGGA GCAGGCTGGG
2401 GCCAGCCCCC ACCGGGTGCA TTACGACCTC AAGTATGCGC TCGGGCTCTG
2451 CGCCGAGCAT GGCCACCACC GCGCTTGTGT CCATGTCTAC AAGGTCTAG
2501 AGCTGTATGA GGAGGCCGTG GACCTGGCCC TGCAGGTGGA TGTGGACCTG
25 2551 GCCAAGCAGT GTGCAGACCT GCCTGAGGAG GATGAGGAAT TGCGAAGAA
2601 GCTGTGGCTG AAGATCGCAC GGCACGTGGT GCAGGAAGAG GAAGATGTAC
2651 AGACAGCCAT GGCTTGCCCTG GCTAGCTGCC CCTTGCTCAA GATTGAGGAT
2701 GTGCTGCCCT TCTTTCCTGA TTTCTCACC ATCGACCACT TCAAGGAGGC
2751 GATCTGCAGC TCACTTAAGG CCTACAACCA CCACATCCAG GCGAGACCTG
30 2801 GGGAGATGGA AGAGGCTACA GCCAGTGCCC AGCGCATCCG GCGAGACCTG
2851 CAGGAGCTGC GGGGCCGCTA CGGCACTGTG GAGCCCCAGG ACAAATGTGC
2901 CACCTGCGAC TTCCCCCTGC TCAACCGCCC TTTTACCTC TTCCTCTGTG
2951 GCCATATGTT CCATGCTGAC TGCCTGCTGC AGGCTGTGCG ACCTGGCCTG
3001 CCAGCCTACA AGCAGGCCCG GCTGGAGGAG CTGCAGAGGA AGCTGGGGGC
35 3051 TGCTCCACCC CCAGCCAAGG GCTCTGCCCG GGCCAAGGAG GCCGAGGGTG
3101 GGGCTGCCAC GGCAGGGCCC AGCCGGGAAC AGCTCAAGGC TGACCTGGAT
3151 GAGTTGGTGG CCGCTGAGTG TGTGTACTGT GGGGAGCTGA TGATCCGCTC
3201 TATCGACCGG CCGTTCATCG ACCCCAGCG CTACGAGGAG GAGCAGCTCA
3251 GTTGGCTGTA GGAGGGTGTG ACCTTTGATG GGGGTGGGCA ATGGGGAGCA
40 3301 GTGGCTTGAA CCCACTTGAG AAGGCTGCCT CCTAGGCTCT GCTCAGTCAT
3351 CTTGCAATTG CCACACTGTG ACCACGTTGA CGGGAGTAGA GTAGCGCTGT
3401 TGGCCAGGAG GTGTCAGGTG TGAGTGTATT CTGCCAGCTT TTCATGCTGT
3451 TCTTCAGAGC TGCAGTTATG CCAGACCATC AGCCTGCCTC CCAGTAGAGG
3501 CCCTTCACCT GGAGAAGTCA GAAATCTGAC CCAATTCCAC CCCCTGCCTC
45 3551 TAGCACCTCT TCTGTCCCTG TCATTCCCCA CACACGTCCT GTTACCTCG
3601 AGAGAGAGAG AGAGAGAGCA CCTTCTTCC GTCTGTTTAC TCTGCGGCCT
3651 CTGGAATCCC AGCTCTTCTC TCTCAGAAGA AGCCTTCTCT TCCTCCTGCC
3701 TGTAGGTGTC CCAGAAGTGA GAAGGCAGCC TTCGAAGTCC TGGGCATTGG
3751 GTGAGAAAGT GATGCTAGTT GGGGCATGCT TTTGTGCACA CTCTCTGGGG
50 3801 CTCCAGTGTG AAGGGTGCCC TGGGGCTGAG GGCCTTGTGG AGGATGGTGG
3851 GTGGTGGTGA TGGAGGTGGA GAGCATTAAG CTGTCTGCAC TGCAAAAAAA
3901 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAGAAAAAA AAAAAAAAAA
3951 A

55

BLAST Results

No BLAST result

Medline entries

- 5 -----
- 97218037:
 Shestopal SA, Makunin IV, Belyaeva ES, Ashburner M, Zhimulev
 IF.;Mol
 10 Gen Genet 1997 Feb 20;253(5):642-8
- 92049306:
 Robinson JS, Graham TR, Emr SD.; A putative zinc finger protein,
 Saccharomyces cerevisiae
 15 Vps18p, affects late Golgi functions required for
 vacuolar protein sorting and efficient alpha-factor
 prohormone maturation. Mol Cell Biol 1991 Dec;11(12):5813-24
- 92049305:
 20 Preston RA, Manolson MF, Becherer K, Weidenhammer E, Kirkpatrick
 D,
 Wright R,
 Jones EW.; Isolation and characterization of PEP3, a gene
 required
 25 for vacuolar biogenesis in Saccharomyces cerevisiae. Mol Cell
 Biol 1991
 Dec;11(12):5801-12

30

Peptide information for frame 1

- 35 ORF from 340 bp to 3258 bp; peptide length: 973
 Category: similarity to known protein
 Classification: Cellular transport and traffic

40 1 MASILDEYEN SLRSRAVLQP GCPSVGIPHS GYVNAQLEKE VPIFTKQRID
 51 FTPSERITSL VVSSNQLCMS LGKDTLLRID LGKANEPNHV ELGRKDDAKV
 101 HKMFLDHTGS HLLIALSSTE VLYVNRNGQK VRPLARWKGQ LVESVGWKA
 151 LGTESSTGPI LVGTAQGHIF EAELSASEGG LFGPAPDLYF RPLYVLNEEG
 201 GPAPVCSLEA ERGPDGRSFV IATTRQRLFQ FIGRAAEGAE AQGFSGLFAA
 251 YTDHPPPFRE FPSNLGYSEL AFYTPKL RSA PRAFAWMMGD GVLYGALDCG
 45 301 RPD SLLSEER VWEYPEGVGP GASPLAIVL TQFHFLLLLA DRVEAVCTLT
 351 GQVVL RDHFL EKFGPLKHMV KDSSTGQLWA YTERAVFRYH VQREARDVWR
 401 TYLDMNRFDL AKEYCRERPD CLDTVLAREA DFCFRQRRYL ESARCYALTQ
 451 SYFEEIALKF LEARQEEALA EFLQRKLASL KPAERTQATL LTTWLTLEYL
 501 SRLGALQGD P EALTLYRETK ECFRTFLSSP RHKEWLFASR ASIHELLASH
 50 551 G DTEH MVYFA VIMQDYERVV AYHCQHEAYE EALAVLARHR DPQLFYKFSP
 601 ILIRHIPRQL VDAWIEMGSR LDARQLIPAL VNYSQGGEVQ QVSQAIRYME
 651 FCVNVLGETE QAIHNYLLSL YARGRPDSLL AYLEQAGASP HRVHYDLKYA
 701 LRLCAEHGHH RACVHVYKVL ELYEEAVDLA LQVDVDLAKQ CADLPEEDEE
 751 LRKKLWLKIA RHVVQEEEDV QTAMACLASC PLLKIEDVLP FFPDFVTIDH
 55 801 FKEAICSSLK AYNHHIQELQ REMEEATASA QRIRDLQEL RGRYGTVEPQ
 851 DKCATCDFPL LNRPFYFLFC GHMFHADCLL QAVRPGLPAY KQARLEELQR
 901 KLGAAPPAK GSARAKEAEG GAATAGPSRE QLKADLDELV AAECVYCCEL
 951 MIRSIDRPFI DPQRYEEEQ L SWL

BLASTP hits

5

No BLASTP hits available

Alert BLASTP hits for DKFZphm12_7g14, frame 1

10 SWISSPROT:DOR_DROME DEEP ORANGE PROTEIN., N = 1, Score = 1279, P
= 2.4e-130

15 PIR:A41943 vacuolar membrane protein PEP3 - yeast (Saccharomyces
cerevisiae), N = 3, Score = 266, P = 5.1e-27

>SWISSPROT:DOR_DROME DEEP ORANGE PROTEIN.
Length = 1,002

20

HSPs:

Score = 1279 (191.9 bits), Expect = 2.4e-130, P = 2.4e-130
Identities = 303/847 (35%), Positives = 463/847 (54%)

25

Query: 130
KVRPLARWKGQLVESVGWUNKALGTESSTGPILVGTAQGHIFEALSASEGGFLFGPAPDLY 189
KVR + ++K + +V +N G ESSTGPIL+GT++G IFE EL+ + G

30 Sbjct: 155 KVRRIEKFKDHEITAVAFNPYHGNESSSTGPILLGTSRGLIFETELNPAADG-
-----HVQ 208

Query: 190 FRPLYVLNEEGGPA-PVCSLEAERGPDG-
RSFVIATTRQRLFAFIGRAAEGAEAGGFSGL 247

35 + LY L G P P+ L+ R P+ R ++ T+ + ++ F +
AE + +

Sbjct: 209 RKQLYDLGL-GRPKYPITGLKLLRVPNSSRYIIVVTSPECIYTF--
QETLKAERSLQAI 265

40 Query: 248 FAAYTD--
HPPPFREFPSNLGYSELAFYTPKLRSAPRAFAWMMGDGVLYGAL--DCGRPD 303
FA Y P E ++L +S+L F+ P P+ +AW+ G+G+ G L

+
Sbjct: 266

45 FAGYVSGVQEPHCEERKTDLTFSQLRFFAPPNSKYPKQWAWLCGEGIRVGELSIEANSAA 325

Query: 304 SLLSEERV---WEYPEGVGPGA---
SPPLAIVLTQFHFLLLADRVEAVCTLTGQVVLRD 357

50 +L+ + +E + G + P A VLT++H +LL AD V A+C L
+ V ++

Sbjct: 326
TLIGNTLINLDFEKTMLHSYGERRLNTPKAFVLTEYHAVLLYADHVRAICLLNQEQVYQE 385

Query: 358 HFLE-
55 KFGPLKHMVKDSSTGQLWAYTERAVFRYHVQREARDVWRTYLDMMNRFDLAKEYCR 416
F E + G + +D TG ++ YT + VF V RE R+VWR YLD
+++LA +

Sbjct: 386

AFDEARVGKPLSIERDEL TGSIIYVYTVKTVFNLRV TREERNVWRIYLDKGQYELATAHAA 445

Query: 417

5 ERPDCLDTVLAREADFCFRQRRYLESARCYALTQSYFEEIALKFLEARQEEALAEFLQRK 476
E P+ L VL + AD F Y +A YA T FEE+ LKF+ +

+ +++++

Sbjct: 446

10 EDPEHLQVLVCQRADAAAFADGSYQVAADYYAETDKSFEEVCLKFMVLPDKRPIINYVKKR 505

Query: 477 LASL--KPAERXXXXXXXXXXXXXXXXXSRLGALQ----

GDPEALTLYRETKEC-FRTFLSS 529

L+ + KP E L L P+ +R + +

+ F+

15 Sbjct: 506

LSRVTTKPMETDELEDKMNIIKALVIWLIDLYLIQINMPDKDEEWRSSWQTEYDEFMME 565

Query: 530

20 PRHKEWLFASRASIELLASHGDTHEMVYFAVIMQDYERVVAYHCQHEAYEEALAVLARH 589
+R ++ +L+A H D +M FA+ + DY+ VVA + E Y

EAL L

Sbjct: 566

AHVLSCTRQNRQTVRQLIAEHADPRNMAQFAIAIGDYDEVVAQQLKAECYAEALQTLINQ 625

25 Query: 590

RDPQLFYKFSPILIRHIPRQLVDWIEMGSRLDARQLIPALVNYSQGGEVQQVSQAIRYM 649
R+P+LFYK++P LI +P+ VDA + GSRL+ +L+P L+ + E ++

+Q RY+

30 Sbjct: 626 RNPELFYKYAPELITRLPKPTVDALMAQGSRLVEKLVPTLI-
IMENREQREQTQ--RYL 682

Query: 650

35 EFCVNVLGETEQAIHNYLLSLYARGRPDSLLAYLEQAGASPHRVHYDLKYALRLCAEHGH 709
EF + L T AIHN+LL LYA P L+ YLE G VHYD+ YA

++C +

Sbjct: 683

EFAIYKLNNTTNDAIHNFLHLHYAEHEPKLLMKYLEIQGRDES LVHYDIYYAHKVCTDL DV 742

Query: 710

40 HRACVHVYKVLLEYEEAVDLALQVDVDLAKQCADLPEEDEELRKKLWLKIARHVVDQEEED 769
A V + +L + AVDLAL D+ LAK+ A P D ++R+KLWL+IA

H ++ D

Sbjct: 743 KEARVFLECMRKWISAVDLALT FDMKLAKETASRPS-

DSKIRRLWLRIAYHDIKGTND 801

45 Query: 770

VQAMACLASCPLLKIEDVLPFFPDVFTIDHFKEAICSSLKAYNHHIQELQREMEETAS 829
V+ A+ L C LL+IED+LPFF DF ID+FKEAIC +L+ YN

IQELQREM E T

50 Sbjct: 802

VKKALNLLKECDLLRIEDLLPFFADFEKIDNFKEAICDALRDYNQRIQELQREMAETTEQ 861

Query: 830

55 AQRIIRDQLQELRGYGTVEPQDKCATCDFPLLNRPFYFLCGHMFHADCLLQAVRPG LPA 889
R +LQ+LR TVE QD C C+ LL +PF++F+CGH FH+DCL +

V P L

Sbjct: 862

TDRATAELQQLRQHSLTVESQDTCEICEMMLLVKPFIFICGHKFHSDCLEKHVVPLLTK 921

Query: 890
YKQARLEELQRKLGAPPXXXXXXXPSREQLKADLDELVAECVYCGE 949
+ RL L+++L A R LK

5 +++++AA+C++CG
Sbjct: 922
EQCRRGLTKRQLEAEVQTQAQPSGALSKQAMELQRKRAALKTEIEDILAADCLFCG- 980

Query: 950 LMIRSIDRPFIDPQRYEEEEQLSW 972
L+I +ID+PF+D +E+ + W
10 Sbjct: 981 LLISTIDQPFVDD--WEQVNVEW 1001

Score = 268 (40.2 bits), Expect = 3.6e-19, P = 3.6e-19
Identities = 91/281 (32%), Positives = 146/281 (51%)

15 Query: 36 QLEKEVPITKQRIDF-TPSE---RITSLVSSNQLCMSLG---
KDTLLRIDLGKANEPN 88
+ ++E IF++ ++ PS + L VS N L LG + TLLR
L +A P

20 Sbjct: 37
ETDEEDEIFSRHKMVLRVPSNCTGDLMLHAVSRNWLVCLLGTPERTTLLRFFLPRAIPPG 96

Query: 89 HVELGRK---DDAKVHKMFLDHTGSHLLIAL---SST-----EVLYVN--
RNGQ----KV 131
25 L + K+ +MFLD TG H++IAL S+T + LY++ +

Q KV
Sbjct: 97
EAVLEKYLSGSGYKITRMFLDPTGHHIIIALVPKSATAGVSPDFLYIHCLESPQAQQLKV 156

30 Query: 132
RPLARWKGQLVESVGWINKALGTESSTGPILVGTAQGHIFEAELSASEGGFLFGPAPDLYFR 191
R + ++K + +V +N G ESSTGPIL+GT++G IFE EL+ + G

+ +
35 Sbjct: 157 RRIEKFKDHEITAVAFNPYHGNESTGPILLGTSRGLIFETELNPAADG---
---HVQRK 210

Query: 192 PLYVLNEEGGPA-PVCSLEAERGPDG-
RSFVIATTRQRLRFQFIGRAAEGAEAGGFSGLFA 249
LY L G P P+ L+ R P+ R ++ T+ + ++ F + AE

40 + +FA
Sbjct: 211 QLYDLGL-GRPKYPITGLKLLRVPNSSRYIIVVTSPECIYTF--
QETLKAERSLQAIFA 267

45 Query: 250 AYTD--HPPPFREFPSNLGYSELAFTPKLRSAPRAFAWMMGDGVLYGAL
297

Y P E ++L +S+L F+ P P+ +AW+ G+G+ G L
Sbjct: 268 GYVSGVQEPHCEERKTDLTFSQLRFFAPPNSKYPKQWAWLCGEGIRVGEL
317

50 Pedant information for DKFZphm12_7g14.1

Report for DKFZphm12_7g14.1

55

[LENGTH] 973
[MW] 110186.09

[pI] 5.72
 [HOMOL] SWISSPROT:DOR_DROME DEEP ORANGE PROTEIN. 1e-145
 [FUNCAT] 30.25 vacuolar and lysosomal organization [S.
 cerevisiae, YLR148w] 5e-41
 5 [FUNCAT] 06.04 protein targeting, sorting and translocation
 [S. cerevisiae, YLR148w] 5e-41
 [FUNCAT] 08.07 vesicular transport (golgi network, etc.) [S.
 cerevisiae, YLR148w] 5e-41
 10 [BLOCKS] BLO0106F Galactokinase proteins
 [BLOCKS] PRO1094B
 [BLOCKS] BPO3306B
 [BLOCKS] PFO0600B
 [PIRKW] yeast vacuole 1e-39
 [PIRKW] transmembrane protein 1e-39
 15 [KW] Alpha_Beta
 [KW] LOW_COMPLEXITY 3.39 %
 [KW] COILED_COIL 4.83 %

20 SEQ MASILDEYENSLRS AVLQPGCP SVGIP HSGYVNAQL EKEVPIFTKQ RIDFTP SERITSL
 SEG
 PRD cccccccccccccccccccccccccccccccccchhhhhhhhhhhhhhhhhhhccccccceeee
 COILS
 25 SEQ VVSSNQLCMSLGKDTLLRIDLG KANEPNHVELGRKDDAKVHKMFLDHTGSHLLIALSSTE
 SEG
 PRD eccccccccccccccccccccccccccccccccceehhhhhhhheeeccccccccccccccccce
 COILS
 30 SEQ VLYVNRNGQKVRPLARWKGQLVESVGWUNKALGTESSTGPILVGTAQGHIFEAELSASEGG
 SEG
 PRD eeeeeccccccchhhhhccchhhhhhhhhhhccc
 35 COILS
 SEQ LFGPAPDLYFRPLYVLNEEGGPAPVCSLEAERGP DGRSFVIATTRQRL FQFIGRAAEAE
 SEG
 40 PRD cchhhhhhhhhhhcchhhhh
 COILS
 SEQ AQGFSGLF AAYTDHPPPFREFPSNLGYSELAFYTPKLRSAPRAFAWMMGDGVLYGALDCG
 SEG
 45 PRD hhhchhhhhhhhhccccccccccccccccccccccccccccccccccccchhhhhhhhhcccccccccccc
 COILS
 50 SEQ RPD SLLSEERVWEYPEGVGP GASPLAIVLTQFHFL LLLADRVEAVCTLTGQVVLRDHFL
 SEG
 PRD cccccchhhhhhhccchhhhhhhhhhhhhhhhh
 55 COILS
 SEQ EKFGPLKHMVKDSSTGQLWAYTERAVFRYHVQREARDVWRTYLD MNRFDLAKEYCRERPD
 SEG
 PRD hccchhhhhhhhhhhhhhhhhccc

[illegible][illegible]

30 SEQ QAIHNYLLSLYARGRPDSSLAYLEQAGASPHRVHYDLKYALRLCAEHGHHRACVHVYKVL
SEG
PRD hhhhhhhhhhhhhccchhhhhhhccccccccchhhhhhhhhhhccccceehhhh
COILS

35 SEQ ELYEEAVDLALQVDVDLAKQCADLPEEDEELRKKLWLKIARHVVQEEEDVQTAMACLASC
SEG
PRD hhhhhhhhhhhhhhhchhhhhhhhhccccchhhhhhhhhhhhhhhhhhhcchhhhhhhhhhhhhc
COILS

[illegible]

.....CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

```

SEQ  RGRYGTVEPQDKCATCDFPLLNRPFYFLCGHMFHADCLLQAVRPGLPAYKQARLEELQR
SEG  .....
PRD  hhhheeeecccccccccccccceeeeeeccchhhhhhhhhhhccchhhhhhhhhhhhh
COILS

```

50 CCCCCCCCC.....

```

SEQ    KLGAAPPAKGSARAKEAEGGAATAGPSREQLKADLDELVAECVYCGELMIRSIDRPF
SEG    .....xxxxxxxxxxxxxxxxxxxxxx.....
PRD    hhhhhcchhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhcccccceeeecccc
55 COTLS

```

.....

SEQ DPQRYEEELSWL

SEG
PRD chhhhhhhhhccc
COILS

5

(No Prosite data available for DKFZphm12_7g14.1)

(No Pfam data available for DKFZphm12_7g14.1)

5 group: melanoma derived

DKFZphmel2_7k19 encodes a novel 234 amino acid protein without similarity to known proteins.

10 Transcripts can be found in almost any tissue, but are most abundant in kidney and retina.
No informative BLAST results; No predictive prosite, pfam or SCOP motife.

15 The new protein can find application in studying the expression profile of melanoma-specific genes.

unknown protein

20 first ATG in frame 1

Sequenced by MediGenomix

25 Locus: /map="3"

Insert length: 2386 bp

Poly A stretch at pos. 2343, polyadenylation signal at pos. 2323

```
30      1 GGCAAAAGTC CAGGAATTAT CTTCATCCCT GGCTATCTTT CTTATATGAA
      51 TGGTACAAAA GCGTTGGCGA TTGAGGAGTT TTGCAAATCT CTAGGTCACG
     101 CCTGCATAAG GTTTGATTAC TCAGGAGTTG GAAGTTCAGA TGGTAACTCA
     151 GAGGAAAGCA CACTGGGGAA ATGGAGAAAA GATGTTCTTT CTATAATTGA
     35 201 TGA CTTAGCT GATGGGCCAC AGATTCTTGT TGGATCTAGC CTTGGAGGGT
     251 GGCTTATGCT TCATGCTGCA ATTGACACGAC CAGAGAAGGT TGTGGCTCTT
     301 ATTGGTGTAG CTACAGCTGC AGATACCTTA GTGACAAAGT TTAATCAGCT
     351 TCCTGTTGAG CTAAAAAAGG AAGTAGAGAT GAAAGGTGTG TGGAGCATGC
     401 CATCAAAATA CTCTGAAGAA GGAGTTTATA ACGTTCAGTA CAGTTTCATT
     45 451 AAAGAAGCTG AACATCACTG CTTGTTACAT AGCCCAATTC CTGTGAACTG
     501 CCCCATAGA TTGCTCCATG GCATGAAGGA TGACATTGTA CCTTGGCATA
     551 CATCAATGCA GGTTGCCGAT CGAGTACTCA GCACAGATGT GGATGTCATC
     601 CTCCGAAAAC ACAGTGATCA CCGAATGAGG GAAAAAGCAG ACATTCAACT
     651 TCTTGTTTAC ACTATTGATG ACTTAATTGA TAAGCTCTCA ACTATAGTTA
     701 ACTAGTATCA CATGTTTAGT TGGTATGTAA ACTAATGTAT CCAGAAGATT
     751 GGAAGAGGGA TAAGAAATGA AAGATCCTGA TACTTTAGGT TTTTCCCTTT
     801 CCTCTATTTT GTAAATATAA GATGAGTATT ATTTAATGAT GTATTTGCAT
     851 AAGTAATGCA AATTGTGAAG AAGGACCAGC TGCTGTTTAG AAAATTTTCT
     901 CCTTCCTTCT GTCCTTGATT TTTTTTCATT AAAGTATTTT CTTTTTTTAA
     50 951 TTCAAGAAAA GTTTACCTTT CTTATGCTTA TGTTAGCTAT GCCAGCTCTT
    1001 AATTGCATCC TTTTCTAATT AGGATTATTA ATAAAGCGTG AATATTTTGT
    1051 TTTTTATTAT AGACAGAAAT TTGTAACATT ACTTCTGATT TGAAGATGCA
    1101 ATTCACAAAA TATAGGGAAA TTTTATTGTA AGTAAATTTG AAATGATGGA
    1151 GAAATTTTCA GAGCATAATA AAGTTCACAA TAAGGATAAT ACTTTATATA
     55 1201 ATGTATAAAG TATATATAAT ATAATATATA TGTTATATAA ACTGCACATT
    1251 ATATTCAAAC TTAAAATTGA GCTTTTTTTT TAAAGGCCCA AAATTGTACA
    1301 GTGATACAAG GAGCTATTTT TAAAATTTGG CTTATGTATA ATATATTTAA
    1351 ATGGGGAATT TCATCTAAAA CAATGATGTA GTATTTTTAA TATTCTGATT
```

1401 GGTAAAATTA AAGAGGAAAT TAATCTTTAT ATATTATTTT TTGCAGAAAC
1451 ATTCATTATT TTATTAATAT TGCCCTAAGT ACAACTAGGC AAGTGATTGC
1501 CACCTAAATC AGAAGACGTT CTAAAGTCAG TAAGAAAGTG TGAAATGCTA
1551 GTATAAAGGT TATTTTTTTT CTTTCCTAAA TAACTAAAGT GAGGTGTAGA
5 1601 TTGAGCCTTG ATATTATTTA GTTAATGTTT TTTATTAATT AATTTTGGCT
1651 GGACTTTATT TAGCTTGATT AGGTTATTAT CTGTCAAACC TTTTAAGTTG
1701 ACAACATGAC TCATATATAT ACATGTGTAT AAGATGAGCA TGTGTCGAAG
1751 ACTTATTCGA CTCATTAATG AGGAAACCAG CAGATAGTAA ACCTGGTTCA
1801 AAGTACAATT CAAGAAACTG AGTATTTATG GGCATTGAAG AAAAAATGTT
10 1851 GAGATAAAAT TGCTGTGCAG AAAAAAGTGT TAATGAAGCC GACCTGACTA
1901 CTTAACCTTA GAGACCTGCT TTACAAGGTT GGCCCTTGAT TGGCATCTGG
1951 GAACCTGGAG TTCAGGGGGC TTCCACCATT CCCAGAACTG ATCAAAGTAG
2001 CTTACTATAT CTAAACTGTA AAACAATATA GTTTCTCCTG AACACCTGCT
2051 TTCCTTCTGG GAGTCTGGAA TTTTGGTATG TGCCAGGCAG AGACTACCTT
15 2101 TGTGACCAGC TCCAGTAAA AACCCAGGC ACTCAGTCTC TAACAAGCTT
2151 TTCTGGTTGA CAGTGTTCCT CAAGTGCTGT TACAAGTGGT TGCTGGGAGA
2201 ATTAAGCTCA TCCTCTGTGA TTCCACTGGC GGAGGATTCT TGGGAAGCTTG
2251 CACTTAGTTT CCCCTGACTT CACCCCATGT GTCTTTTTTC CTTTGCTGAT
2301 TTTGTTTTGT ATCCTTTCAC TGTAATAAAT CATGGCCGTG AGCAGAAAAA
20 2351 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAA

BLAST Results

25

No BLAST result

Medline entries

30

No Medline entry

35

Peptide information for frame 1

ORF from 46 bp to 702 bp; peptide length: 219

40

Category: similarity to unknown protein

Classification: unclassified

45

1 MNGTKALAIE EFCKSLGHAC IRFDYSGVGS SDGNSEESTL GKWRKDVLSI
51 IDDLADGPGI LVGSSLGGWL MLHAAIARPE KVALIGVAT AADTLVTKFN
101 QLPVELKKEV EMKGVWSMPS KYSEEGVYNV QYSFIKEAEH HCLLHSPPIV
151 NCPILLHGM KDDIVPWHTS MQVADRVLST DVDVILRKHS DHRMREKADI
201 QLLVYTIDDL IDKLSTIVN

50

BLASTP hits

No BLASTP hits available

55

Alert BLASTP hits for DKFZphm12_7k19, frame 1

No Alert BLASTP hits found

Pedant information for DKFZphmel2_7k19, frame 1

Report for DKFZphmel2_7k19.1

5

[LENGTH] 219
 [MW] 24309.18
 [pI] 5.69
 [HOMOL] PIR:A71691 hypothetical protein RP343 - Rickettsia
 prowazekii 3e-29
 [BLOCKS] BP04352K
 [BLOCKS] PR00828E
 [KW] Alpha_Beta

10

15

SEQ MNGTKALAIIEEFCKSLGHACIRFDYSGVGSSDGNSEESTLGKWRKDVLSIIDDLADGPQI
 PRD ccchhhhhhhhhhhhhccceeeeeeeccccccccccccccccchhhhhhhhhhhhhccceee

20

SEQ LVGSSLGGWMLHAAIARPEKVVALIGVATAADTLVTKFNQLPVELKKEVEMKGVWSMPS
 PRD eeccccchhhhhhhhhhhccceeeeeeeeeehhhhhhccccchhhhhhhhhhhhhheeeccc

SEQ KYSEEGVYNVQYSFIKEAEHHCLLHSPIPVNCPIRLLHGMKDDIVPWHTSMQVADRVLST
 PRD cccccceeeehhhhhhhhhhhhhhhccccccccceccccccccccccchhhhhhhhhhhhh

25

SEQ DVDVILRKHSDHRMREKADIQLLVYTIDDLIDKLSTIVN
 PRD hheeeeeccccchhhhhhhheeeeeehhhhhhhhhcccccc

30

(No Prosite data available for DKFZphmel2_7k19.1)

(No Pfam data available for DKFZphmel2_7k19.1)

35

DKFZphtes3_10i16

group: nucleic acid management

40

DKFZphtes3_10i16 encodes a novel 742 amino acid protein with similarity to human ZK1.

45

The ZK1 gene is one of early response genes by exposure to ionizing radiation, and plays a role in radiation-induced apoptotic cell death on hematopoietic cells. The novel protein contains 18 zinc finger domains, a RGD cell attachment and a ATP GTP A domain.

50

The new protein can find application in diagnosis/therapy in leukemia predisposition/disease in the modulation of DNA repair.

55

similarity to ZK1 (Homo sapiens), complete cds.

Sequenced by Qiagen

Locus: unknown

Insert length: 2884 bp

Poly A stretch at pos. 2861, polyadenylation signal at pos. 2835

```
5      1 CGGAAATGGA GGGGGTCTGCT TTCCTCACCT TCCTCGCTGC GCGGGCGGCG
      51 GTTGGTAACC GGTGAGACCA GCCCAGAGAG GACCTGGTGC CTGTACCCAG
     101 GCTTCTGTCT CTCTGTCTGCC TGCCTATATG CCTGCTGTAG TCACAGGAGC
     151 TGTAGAGAGG ACCCCGGTAC ATCTGAAAGC CGGGAAATGG ACCCAGTGGC
     201 CTTTGAGGAT GTGGCTGTGA ACTTCACCCA GGAAGAGTGG ACATTGCTGG
     251 ATATTTCCCA GAAGAATCTC TTCAGGGAAG TGATGCTGGA AACTTTCAGG
     301 AACCTGACCT CTATAGGAAA AAAATGGAGT GACCAGAACA TTGAATATGA
     351 GTACCAAAAC CCCAGAAGAA GCTTCAGGAG TCTCATAGAA GAGAAAGTCA
     401 ATGAAATTAA AGAAGACAGT CATTGTGGAG AAACCTTTAC CCAGGTTCCA
     451 GATGACAGAC TGAACCTTCCA GGAGAAGAAA GCTTCTCCTG AAGTAAATC
     501 ATGTGACAGC TTTGTGTGTG CAGAAGTTGG CATAGGTAAC TCATCTTTTA
     551 ATATGAGCAT CAGAGGTGAC ACTGGACACA AGGCATATGA GTATCAGGAA
     601 TATGGACCAA AGCCATATAA GTGTCAACAA CCTAAAAATA AGAAAGCCTT
     651 CAGGTATCGC CCATCCATTA GAACACAAGA AAGGGATCAC ACTGGAGAGA
     701 AACCTATGCT TTGTAAAGTC TGTGGAAAAA CTTTTATTTT CCATTCAAGC
     751 ATTGGAAGAC ACATGGTAAT GCACAGTGGG GATGGAACCT ATAAATGTAA
     801 ATTTTGTGGG AAAGCCTTCC ATTCTTTTCT TTTATATCTT ATCCATGAAA
     851 GAACTCACAC TGGAGAGAAA CCATATGAAT GTAAACAATG TTGTAATCC
     901 TTTACTTATT CTGCTACCCT TCAAATACAT GAAAGAACTC ACCTGGGGA
    25  951 GAAGCCCTAT GAATGTAGCA AATGTGATAA AGCATTTCAT AGTCTAGTT
    1001 CCTATCATAG ACATGAAAGA AGTCACATGG GAGAGAAGCC TTATCAATGC
    1051 AAAGAATGTG GAAAAGCATT TGCATATACC AGTTCTCTTC GTAGACATGA
    1101 AAGGACCCAC TCTGGGAAAA AACCGTATGA ATGTAAGCAA TATGGGGAAG
    1151 GCTTATCCTA TCTTATAAGT TTTCAAACAC ACATAAGAAT GAACTCTGGA
    30  1201 GAAAGACCTT ATAAATGTAA GATATGTGGG AAAGGCTTTT ATTCTGCCAA
    1251 GTCATTTCAA ACACATGAAA AAACCTCACAC TGGAGAGAAA CCACTGAAAT
    1301 GCAAGCAATG TGGTAAAGCC TTCAATCTTT CCAGTTCCTT TCGATATCAT
    1351 GAAAGGATTC ACACTGGAGA GAAACCCTAT GAGTGTAAGC AGTGTGGGAA
    1401 AGCCTTCAGA TCTGCCTCAC AGCTTCGAGT GCACGGTGGG ACTCACACTG
    35  1451 GAGAGAAACC CTATGAATGT AAGGAATGTG GGAAAGCCTT CAGATCTACC
    1501 TCACACCTTC GAGTGCATGG TAGGACTCAT ACTGGAGAGA AACCTATGA
    1551 ATGTAAGGAA TGTGGGAAAG CCTTCAGATA TGTGAAGCAC CTTCAAATTC
    1601 ATGAAAGGAC AGAAAAACAC ATAAGAATGC CCTCTGGAGA AAGACCTTAT
    1651 AAATGTAGTA TATGTGAGAA AGGCTTTTAT TCTGCCAAGT CATTCAATAC
    40  1701 ACATGAAAAA ACTCACACTG GAGAGAAACC CTATGAATGC AACCAATGTG
    1751 GTAAAGCCTT CAGATGTTGC AATTCCCTTC GATATCATGA AAGGACTCAC
    1801 ACTGGAGAGA AACCTATGA GTGTAAGCAA TGTGGGAAAG CCTTCAGATC
    1851 TGCCTCACAC CTTGGAATGC ATGAAAGGAC TCACACTGGA GAGAAACCCT
    1901 ATGAGTGTA GCAATGTGGG AAAGCCTTCA GTTGTGCCTC AAACCTTCGA
    45  1951 AAGCATGGTA GGAATCACAC TGGAGAGAAA CCCTATGAGT GTAAGCAATG
    2001 TGGGAAAGCC TTCAGATCTG CCTCAAACCT TCAGATGCAT GAAAGGACTC
    2051 ACACTGGAGA GAAACCCTAT GAATGTAAGG AATGCGAAAA AGCATTCTGT
    2101 AAATTCTCTT CTTTTCAAAT ACATGAAAGG AAGCACAGAG GAGAGAAGCC
    2151 CTATGAATGT AAGCATTGTG GGAATGGATT CACATCTGCC AAGATTCTTC
    50  2201 AAATACATGC AAGAACACAC ATTGGAGAGA AACACTATGA ATGTAAGGAA
    2251 TGCAGGAAAAG CATTCAATTA TTTTCTTCC TTGCATATAC ACGCAAGGAC
    2301 TCATATGGGA GAGAAGCCAT ATGAATGTAA GGATTGTGGG AAAGCATTCA
    2351 GCTAGCCTGG TTCCTTTTAT GGACATGAAT AGACTCACAC TGGAAAGGAAG
    2401 CACTATGAAT GCAAGCAATG TGGCAAAACT TTCACATTTT CCAGTTCTTT
    55  2451 TCGATATCAT GAAAGGACTC ACACTGGGGA GAAACCCTAT CAATGTAAGC
    2501 AGTGTGGGAA AGCCTTCATT CCTTTTACTT CTTTTCATATG TCATGAAAGG
    2551 ACTCACACGG GAGAGAAACC CTATGAGTGT ATTCTAGTTC CGTTTGATAT
    2601 CATGAAAGGA CTTACACTGG AGTGAAACCC TATGAATGTA AGCAATGTGG
```

2651 GAAAGCCTTC AGATGTGCCT CGCACCTTCA ACGGCATGGA AGGGTTCACA
 2701 CTTGGGAGAA ACTCTATGAA TGTAAGCAGT ATGGGAAAGC CTTGAGATCT
 2751 GCCAAGATTC TTTGAATACA GATAATTAAT GTAAACAATT ATCATAAGTA
 2801 TACTAACATG TTATTCTTTT TAAATAAGAA GGTATAATAA AATATCCCAT
 5 2851 TGGTTTTATG TATTAAAAAA AAAAAAAAAA AAAA

BLAST Results

10

No BLAST result

Medline entries

15

98401134:

Katoh O, Oguri T, Takahashi T, Takai S, Fujiwara Y, Watanabe H.;
 ZK1. a

20 novel Kruppel-type zinc finger gene, is induced following
 exposure to ionizing radiation and enhances apoptotic cell death
 on hematopoietic cells. Biochem Biophys Res Commun 1998 Aug
 28;249(3):595-600

25 95137393:

Wick MJ, Ann DK, Lee NM, Loh HH.; Isolation of a cDNA encoding a
 novel

zinc-finger protein from

neuroblastoma x glioma NG108-15 cells. Gene 1995 Jan

30 23;152(2):227-32

35

Peptide information for frame 1

ORF from 127 bp to 2352 bp; peptide length: 742

Category: similarity to known protein

40 Classification: Nucleic acid management

Prosite motifs: RGD (146-148)

ATP_GTP_A (195-202)

ZINC_FINGER_C2H2 (196-216)

ZINC_FINGER_C2H2 (224-244)

45 ZINC_FINGER_C2H2 (252-272)

ZINC_FINGER_C2H2 (280-300)

ZINC_FINGER_C2H2 (308-328)

ZINC_FINGER_C2H2 (364-384)

ZINC_FINGER_C2H2 (392-412)

50 ZINC_FINGER_C2H2 (420-440)

ZINC_FINGER_C2H2 (448-468)

ZINC_FINGER_C2H2 (510-530)

ZINC_FINGER_C2H2 (538-558)

ZINC_FINGER_C2H2 (566-586)

55 ZINC_FINGER_C2H2 (594-614)

ZINC_FINGER_C2H2 (622-642)

ZINC_FINGER_C2H2 (650-670)

ZINC_FINGER_C2H2 (678-698)

ZINC_FINGER_C2H2 (706-726)
ZINC_FINGER_C2H2 (476-498)

```

5      1 MPCCSHRSCR EDPGTSESRE MDPVAFEDVA VNFTQEEWTL LDISQKNLFR
      51 EVMLETFRNL TSIGKKWSDQ NIEYEYQNP RSFRSLIEEK VNEIKEDSHC
     101 GETFTQVPDD RLNFQEKKAS PEVKSCDSFV CAEVGIGNSS FNMSIRGDTG
     151 HKAYEYQYEG PKPYKCQQPK NKKAFRYRPS IRTQERDHTG EKPYPACKVCG
     201 KTFIFHSSIR RHMVMHSGDG TYKCKFCGKA FHSFSLYLIH ERTHTGEKPY
    10 251 ECKQCGKSFT YSATLQIHER THTGEKPYEC SKCDKAFHSS SSYHRHERSH
     301 MGEKPYQCKE CGKAFAYTSS LRRHERTHSG KKPYECKQYG EGLSYLISFQ
     351 THIRMNSGER PYKCKICGKG FYSAKSFQTH EKTHTGEKRY KCKQCGKAFN
     401 LSSSFYHER IHTGEKPYEC KQCGKAFRSA SGLRVHGGTH TGEKPYECKE
     451 CGKAFRSTSH LRVHGRTHTG EKPYECKECG KAFRYVKHLQ IHERTEKHIR
    15 501 MPSGERPYKC SICEKGFYSA KSFQTHEKTH TGEKPYECNQ CGKAFRCCNS
     551 LRYHERHTTG EKPYECKQCG KAFRSASHLR MHERHTTGEK PYECKQCGKA
     601 FSCASNLRKH GRHTTGEKPY ECKQCGKAFR SASNLQMHHER THTGEKPYEC
     651 KECEKAFCKF SSFQIHERKH RGEKPYECKH CGNGFTSAKI LQIHARTHIG
     701 EKHYECKECG KAFNYFSSLH IHARTHMG EK PYECKDCGKA FS
20

```

BLASTP hits

25 No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_10i1b, frame 1

30 No Alert BLASTP hits found

Peptide information for frame 2

35 ORF from 1703 bp to 2584 bp; peptide length: 294
Category: questionable ORF
Classification: no clue

```

40      1 MKKLTLERNP MNATNVVKPS DVAIPFDIMK GLTLERNPMS VSNVGKPSDL
      51 PHTFECMKGL TLERNPMSVS NVGKPSVVPQ TFESMVGLTL ERNPMSVSNV
     101 GKPSDLPQTF RCMKGLTLER NPMNVRNAK HSVNSLLFKY MKGSTEERSP
     151 MNVSIVGMDS HLPRFFKYMQ EHTLERNTMN VRNAEKHSII FLPCIYTQGL
     201 IWERSHMNVR IVGKHSASLV PFMDMNRLLT EGSTMNASNV AKLSHFVPLF
     251 DIMKGLTLGR NPINVSSVGK PSFLLLLFNV MKGLTRERNP MSVF
45

```

BLASTP hits

50 No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_10i1b, frame 2

55 TREMBL:AF153201_1 product: "zinc finger protein dp"; Homo
sapiens zinc
finger protein dp mRNA, complete cds.; N = 1, Score = 225, P =
4.1e-18

>TREMBL:AF153201_1 product: "zinc finger protein dp"; Homo sapiens zinc finger protein dp mRNA, complete cds.
Length = 423

HSPs:

Score = 225 (33.8 bits), Expect = 4.1e-18, P = 4.1e-18
Identities = 84/246 (34%), Positives = 122/246 (49%)

Query: 16 VVKPSDVA-
IPFDIMKGLTLERNPMSVSNVGKPSDLPHTFECMKGLTLERNPMSVSNVGK 74
V KPS A I F I + + L RN + V +V K S T ++G

15 TLERNP++V +VGK
Sbjct: 3 VGKPSVRAQILFCIRESI-
LGRNHIHVISVAKVSVRIQTLLNIEGSTLERNPINVM SVGK 61

Query: 75
20 PSVVPQT FESMVGLTLERNPMSVSNVGKPSDLPQTFRCKGLTLERNPMNVRNAKKHSVN 134
+ A K V + Q+ + G LERNP+ V NV KPS Q + TLER+ +V

Sbjct: 62
25 LLIRAQSLFYIRGFILERNPIPVINVAKPSVGFQILLIINEFTLERSLTHVISAIAKCLVE 121

Query: 135 SLLFKYMKGSTEERSPMNV SIVGMDS-
HLPRFFKYMQEHTLERNTMNV RNAEKHSIIFLP 193
+ + + R+PMNV VG P F +++E TLERN M+V

30 K +
Sbjct: 122 DEILLNITEFIQVRNPMNVMNVGKPLVRAPTLF-
FIRESTLERNL MHVVIVLKALVAVQI 180

Query: 194
35 CIYTQGLI WERSHMNVRI VGKHSASLVPFMDMNRLTLEGSTMNASNVAKLSHFPVLFDIM 253
+ + ER+HM+V V K +++ TL S + A V K S

Sbjct: 181
LLSIKEYTLERNHMHVISVIKVLVKAQTS LNIREYTLVKSLIIAIVVRKPSVRVLT LFFI 240

40 Query: 254 KGLTLGRN 261
+ TL +N
Sbjct: 241 REFTLEKN 248

Score = 215 (32.3 bits), Expect = 1.1e-16, P = 1.1e-16
45 Identities = 82/246 (33%), Positives = 124/246 (50%)

Query: 44
VGKPSDLPHTFECMKGLTLERNPMSVSNVGKPSVVPQT FESMVGLTLERNPMSVSNVGK 103
VGKPS C++ L RN + V +V K SV QT ++G

50 TLERNP++V +VGK
Sbjct: 3
VGKPSVRAQILFCIRESILGRNHIHVISVAKVSVRIQTLLNIEGSTLERNPINVM SVGKL 62

Query: 104 SDLPQTFRCKGLTLERNPMNVRNAKKHSVNSLLFKYMKGSTEERSPMNV-
55 SIVGM---D 159
Q+ ++G LERNP+ V N K SV + + T ERS +V S
+ D

Sbjct: 63
LIRAQSLFYIRGFILERNPIPVINVAKPSVGFQILLIINEFTLERSLTHVISAIAKCLVED 122

Query: 160 SHLPRFFKYMQEHTLERNTMNVRNAEKHSIIFLPCIY-
TQGLIWEERSHMNVIRIVGKHSAS 218

L +++Q RN MNV N K ++ P ++ + ER+ M+V
IV K +

Sbjct: 123 EILLNITEFIQV----RNPMNVMNVGK-
PLVRAPTLLFFIRESTLERNLMHVIVLKAALVA 177

Query: 219 LVPFMDMNRLTLEGSTMNASNVAK-
LSHFPVLFDIMKGLTLGRNPINVSSVGKPSFLLL 277

+ + + TLE + M+ +V K L +I + TL ++ I V
KPS +L

Sbjct: 178 VQILLSIKEYTLERNMHMVISVIKVLVKAQTSLNIRE-
YTLVKSLIIAIVVRKPSVRVLT 236

Query: 278 FNVMKGLTRERN 289
++ T E+N

Sbjct: 237 LFFIREFTLEKN 248

Score = 207 (31.1 bits), Expect = 5.2e-15, P = 5.2e-15
Identities = 80/270 (29%), Positives = 129/270 (47%)

Query: 1 MKKLTLERNPMTNATNVVKPSDVAIPFDI-
MKGLTLERNPMSVSNVGKPSDLPHTFECMKG 59

+++ L RN ++ +V K S V I + ++G TLERNP++V +VGK
+ ++G

Sbjct: 16 IRESILGRNHIHVISVAKVS-
VRIQTLLNIEGSTLERNPINVMSVGKLLIRAQSLFYIRG 74

Query: 60
LTLERNPMSVSNVGKPSVVPQTFESMVGLTLERNPMSVSNVGKPSDLPQTFRCMKGLTLE 119
LERNP+ V NV KPSV Q + TLER+ V + K +

Sbjct: 75
FILERNPIPVINVAKPSVGFQILLIINEFTLERSLTHVISAIAKCLVEDEILLNITEFIQV 134

Query: 120
RNPMNVRNAKKHSVNSLLFKYMKGSTEERSPMNVSIIVGMDSHLPRFFKYMQEHTLERNTM 179
RNPMNV N K V + +++ ST ER+ M+V IV +

++E+TLERN M
Sbjct: 135
RNPMNVMNVGKPLVRAPTLLFFIRESTLERNLMHVIVLKAALVAVQILLSIKEYTLERNHM 194

Query: 180
NVRNAEKHSIIFLPCIYTTQGLIWEERSHMNVIRIVGKHSASLVPFMDMNRLTLEGSTMNASN 239
+V + K + + + +S + +V K S ++ + TLE

Sbjct: 195
HVISVIKVLVKAQTSLNIREYTLVKSLIIAIVVRKPSVRVLTLLFFIREFTLEKNYYLCTQ 254

Query: 240 VAKLSHFPVLFDIMKGLTL--GRNPINVSSVGK 270
+K F + D++K + G P S K

Sbjct: 255 CSK--SFSQISDLIKHQRHTGEKPYKCSECRK 285

Score = 181 (27.2 bits), Expect = 1.4e-11, P = 1.4e-11
Identities = 74/269 (27%), Positives = 116/269 (43%)

Query: 5
TLERNPMNATNVVKPSDVAIPFDIMKGLTLERNPMNSVSNVGKPSDLPHTFECMKGLTLER 64
TLERNP+N +V K A ++G LERNP+ V NV KPS

5 + TLER
Sbjct: 48
TLERNPINVMSVGKLLIRAQSLFYIRGFILERNPIPVINVAKPSVGFQILLIINEFTLER 107

Query: 65
10 NPMSVSNVGKPSVVPQTFESMVGLTLERNPMNSVSNVGKPSDLPQTFRCMKGLTLERNPMN 124
+ V + K V + ++ RNPM+V NVGKP T ++
TLERN M+
Sbjct: 108
SLTHVISAIAKCLVEDEILLNITEFIQVRNPMNVMNVGKPLVRAPTLFFIRESTLERNLMH 167

15 Query: 125 VRNAKKHSVNSLLFKYMKGSTEERSPMNV-
SIVGMDSHLPRFFKYMQEHTLERNTMNVRN 183
V K V + +K T ER+ M+V S++ + ++E+TL
++ +

20 Sbjct: 168 VVIVLKALVAVQILLSIKEYTLERNHMHVISVIKVLVKAQTSLN-
IREYTLVKSLIIAIV 226

Query: 184
25 AEKHSIIFLPCIYTQGLIWERSHMNVRIVGKHSASLVPFMDMNRLTLEGSTMNASNVAKL 243
K S+ L + + E+++ K + + + R+
S K
Sbjct: 227
VRKPSVRVLTFFIREFTLEKNYYLCTQCSKSFSQISDLIKHQRIHTGEKPYKCSECRKA 286

30 Query: 244 SHFPVLFDIMKGLTLGRNPINVS SVGKPSF 273
L + + + G+ P GK SF
Sbjct: 287 FSQCSLLALHQRIHTGKKPNPCDECGK-SF 315

Score = 166 (24.9 bits), Expect = 8.4e-10, P = 8.4e-10
35 Identities = 63/194 (32%), Positives = 89/194 (45%)

Query: 100
VGKPSDLPQTFRCMKGLTLERNPMNVRNAKKHSVNSLLFKYMKGSTEERSPMNVSIVGMD 159
VGKPS Q C++ L RN ++V + K SV ++GST

40 ER+P+NV VG
Sbjct: 3
VGKPSVRAQILFCIRESILGRNHIHVISVAKVSVRIQTLLNIEGSTLERNPINVMSVGKL 62

Query: 160
45 SHLPRFFKYMQEHTLERNTMNVRNAEKHSIIFLPCIYTQGLIWERSHMNVRIVGKHSASL 219
+ Y++ LERN + V N K S+ F + ERS +V
K
Sbjct: 63
LIRAQSLFYIRGFILERNPIPVINVAKPSVGFQILLIINEFTLERSLTHVISAIAKCLVED 122

50 Query: 220 VPFMDMNRLTLEGSTMNASNVAK-
LSHFPVLFDIMKGLTLGRNPINVS SVGKPSFLLLLF 278
+++ + MN NV K L P LF I + TL RN ++V V K
+ +

55 Sbjct: 123 EILLNITEFIQVRNPMNVMNVGKPLVRAPTLFFIRES-
TLERNLMHVIVLKALVAVQIL 181

Query: 279 NVMKGLTRERNPMSV 293

+K T ERN M V
 Sbjct: 182 LSIKEYTLERNHMHV 196

5 Pedant information for DKFZphtes3_10i1b, frame 1

Report for DKFZphtes3_10i1b.1

10 [LENGTH] 784
 [MW] 90857.05
 [pI] 9.24
 [HOMOL] TREMBL:AB011414_1 gene: "ZK1"; product: "Kruppel-
 15 type zinc finger protein"; Homo sapiens ZK1 mRNA for Kruppel-type
 zinc finger protein, complete cds. 0.0
 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YJL056c] 6e-33
 [FUNCAT] 04.05.01.04 transcriptional control [S. cerevisiae,
 20 YJL056c] 6e-33
 [FUNCAT] 04.99 other transcription activities [S. cerevisiae,
 YOR113w] 5e-24
 [FUNCAT] 04.01.01 rrna synthesis [S. cerevisiae, YPR186c PZF] -
 TFIIIA] 1e-20
 25 [FUNCAT] 04.03.01 trna synthesis [S. cerevisiae, YPR186c PZF] -
 TFIIIA] 1e-20
 [FUNCAT] 13.04 homeostasis of other ions [S. cerevisiae,
 YNL027w] 1e-13
 [FUNCAT] 11.07 detoxification [S. cerevisiae, YGL254w] 2e-12
 30 [FUNCAT] 01.02.04 regulation of nitrogen and sulphur utilization
 [S. cerevisiae, YGL254w] 2e-12
 [FUNCAT] 01.05.04 regulation of carbohydrate utilization [S.
 cerevisiae, YGL209w] 2e-11
 [FUNCAT] 04.05.99 other mrna-transcription activities [S.
 35 cerevisiae, YER028c] 3e-10
 [FUNCAT] 11.01 stress response [S. cerevisiae, YKLO62w] 1e-09
 [FUNCAT] 01.01.04 regulation of amino-acid metabolism [S.
 cerevisiae, YDR253c] 5e-09
 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YBR066c]
 40 3e-08
 [FUNCAT] 03.07 pheromone response, mating-type determination,
 sex-specific proteins [S. cerevisiae, YDR146c] 1e-07
 [FUNCAT] 03.25 cytokinesis [S. cerevisiae, YLR131c] 2e-06
 [BLOCKS] BL00466 TFIIS zinc ribbon domain proteins
 45 [BLOCKS] BL00245A Phytochrome chromophore attachment site
 proteins
 [BLOCKS] DM01951B
 [BLOCKS] PF01363B
 [BLOCKS] BL01030
 50 [BLOCKS] PF00096B
 [BLOCKS] BL00028 Zinc finger, C2H2 type, domain proteins
 [BLOCKS] BP04213E
 [BLOCKS] BP04213C
 [BLOCKS] BP04213B
 55 [SCOP] d2adr_ 7.31.1.1.4 ADR1 [synthetic based on yeast
 (Saccharomyce 2e-05
 [PIRKW] nucleus 1e-53
 [PIRKW] RNA binding 2e-58

[PIRKW] duplication 1e-34
 [PIRKW] tandem repeat 1e-171
 [PIRKW] spermatogenesis 5e-62
 [PIRKW] zinc 1e-169
 5 [PIRKW] zinc finger 0.0
 [PIRKW] DNA binding 0.0
 [PIRKW] metal binding 1e-120
 [PIRKW] phosphoprotein 2e-58
 [PIRKW] leucine zipper 1e-53
 10 [PIRKW] alternative splicing 2e-58
 [PIRKW] eye lens 1e-111
 [PIRKW] oocyte 1e-106
 [PIRKW] transcription factor 1e-111
 [PIRKW] embryo 1e-106
 15 [PIRKW] segmentation 1e-34
 [PIRKW] transcription regulation 1e-152
 [SUPFAM] POZ domain homology 7e-83
 [SUPFAM] transcription factor Krueppel 1e-34
 [SUPFAM] zinc finger protein ZFP-36 1e-173
 20 [SUPFAM] transcription factor IIIA 8e-31
 [PROSITE] ATP_GTP_A 1
 [PROSITE] RGD 1
 [PROSITE] ZINC_FINGER_C2H2 18
 [PFAM] Zinc finger, C2H2 type
 25 [PFAM] TNFR/NGFR cysteine-rich region
 [KW] Irregular
 [KW] 3D
 [KW] LOW_COMPLEXITY 3.57 %
 30
 SEQ RKWRGSLSSPSSLRGRRLVTGQTSPRGTWCLYPGFCRSVACAMPCCSHRSCREDPGTSES
 SEG
 1meyF
 35
 SEQ REMDPVAFEDVAVNFTQEEWTLDDISQKNLFREVMLETFRNLTSIGKKWSDQNIYEYQN
 SEG
 1meyF
 40
 SEQ PRRSFRSLIEEKVNEIKEDSHCGETFTQVPDDRLNFQEKKASPEVKSCDSFVCAEVGIGN
 SEG
 1meyF
 45
 SEQ SSFNMSIRGDTGHKAYEYQ EYGP KP YKCQ QPK NK KAF RYRPSIRTQERDHTGEKPYACKV
 SEG
 1meyF
 50
 SEQ CGKTFIFHSSIRRHMMHSGDGT YKCKFCGKAFHSFSLYLIHERTHTGEKPYECKQCGKS
 SEG
 1meyF
 55
 SEQ FTYSATLQIHERTHTGEKPYECSKCDKAFHSSSSYHRHERSHMGEKPYQCKECCGKAFAYT
 SEG

lmeyF

5 SEQ SSLRRHERTHSGKKPYECKQYGEGLSYLISFQTHIRMNSGERPYKCKICGKGFYSAKSFQ
SEG
lmeyF

10 SEQ THEKTHTGEKRYKCKQCGKAFNLSSSFYHERIHTGEKPYECKQCGKAFRSASQLRVHGG
SEG
lmeyF

15 SEQ THTGEKPYECKECGKAFRSTSHLRVHGRHTHTGEKPYECKECGKAFRYVKHLQIHERTEKH
SEG
lmeyF

20 SEQ IRMPSGERPYKCSICEKGFYSAKSFQTHEKTHTGEKPYECNCGKAFRCCNSLRHERTH
SEG
lmeyF

25 SEQ TGEKPYECKQCGKAFRSASHLRMHERTHTGEKPYECKQCGKAFSCASNLRKHGRHTHTGEK
SEG
lmeyF ..TTTEETTTTCEETTHHHHHHHHHHHHTTCCEEETTTTEEECCHHHHHHHHHHHHCC

30 SEQ PYECKQCGKAFRSASNLMHERHTHTGEKPYECKEKEKAFCKFSSFQIHERKHRGEKPYEC
SEG
lmeyF CEEETTTTEEECCHHHHHHHHHHHH.....

35 SEQ KHCNGFTSAKILQIHARTHIGEKGHYECKECGKAFNYFSSLHIHARTHMGEEKPYECKDCG
SEG
lmeyF

40 SEQ KAFS
SEG
lmeyF

45 Prosite for DKFZphtes3_10116.1

PS00016	188->191	RGD	PD0C00016
PS00017	237->245	ATP_GTP_A	PD0C00017
PS00028	238->259	ZINC_FINGER_C2H2	PD0C00028
50 PS00028	266->287	ZINC_FINGER_C2H2	PD0C00028
PS00028	294->315	ZINC_FINGER_C2H2	PD0C00028
PS00028	322->343	ZINC_FINGER_C2H2	PD0C00028
PS00028	350->371	ZINC_FINGER_C2H2	PD0C00028
PS00028	406->427	ZINC_FINGER_C2H2	PD0C00028
55 PS00028	434->455	ZINC_FINGER_C2H2	PD0C00028
PS00028	462->483	ZINC_FINGER_C2H2	PD0C00028
PS00028	490->511	ZINC_FINGER_C2H2	PD0C00028
PS00028	552->573	ZINC_FINGER_C2H2	PD0C00028

	PS00028	580->601	ZINC_FINGER_C2H2	PD0C00028
	PS00028	608->629	ZINC_FINGER_C2H2	PD0C00028
	PS00028	636->657	ZINC_FINGER_C2H2	PD0C00028
	PS00028	664->685	ZINC_FINGER_C2H2	PD0C00028
5	PS00028	692->713	ZINC_FINGER_C2H2	PD0C00028
	PS00028	720->741	ZINC_FINGER_C2H2	PD0C00028
	PS00028	748->769	ZINC_FINGER_C2H2	PD0C00028
	PS00028	518->541	ZINC_FINGER_C2H2	PD0C00028

10

Pfam for DKFZphtes3_10116.1

15 HMM_NAME TNFR/NGFR cysteine-rich region

HMM *CpeGtYtD-WNHvpqClpC-.trCePEMGQYMvqPCTwTQNTVC*
 C + +++ +++++C C ++C+++ G+++++ ++ V
 Query 30 CLYPGFCRSVACAMPC--CSHRSCREDPGTSESREMDP----VA
 20 67

25 HMM_NAME Zinc finger, C2H2 type

HMM *CpwPDCgKtFrrwsNLrRHMRT*
 C++ CGKTF S+ RRHM +H
 Query 238 CKV--CGKTFIFHSSIRRHVMH 258

30 32.15 (bits) f: 266 t: 286 Target: dkfzphes3_10116.1
 similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

Query *CpwPDCgKtFrrwsNLrRHMRT*
 C++ CGK+F + S + +H RTH

35 dkfzphes3 266 CKF--CGKAFHSFSLYLIHERTH 286

Query f: 294 t: 314 Target: dkfzphes3_10116.1
 similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

40 HMM *CpwPDCgKtFrrwsNLrRHMRT*
 C+ CGK+F+++ +L++H RTH

Query 294 CKQ--CGKSFTYSATLQIHERTH 314

45 34.22 (bits) f: 322 t: 342 Target: dkfzphes3_10116.1
 similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

Query *CpwPDCgKtFrrwsNLrRHMRT*
 C++ C+K+F ++S++ RH R+H

50 dkfzphes3 322 CSK--CDKAFHSSSSYHRHERSH 342

Query f: 350 t: 370 Target: dkfzphes3_10116.1
 similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

55 HMM *CpwPDCgKtFrrwsNLrRHMRT*
 C++ CGK+F + S+LRRH RTH

Query 350 CKE--CGKAFAYTSSLRRHERTH 370

32.09 (bits) f: 406 t: 426 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

Query *CpwPDGgKtFrrwsNLrRHMRTTH*

C++ CGK F ++ ++++H +TH

dkfzphtes3 406 CKI--CGKGFYSAKSFQTHEKTH 426

Query f: 434 t: 454 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

HMM *CpwPDGgKtFrrwsNLrRHMRTTH*

C+ CGK+F+ +S++R H R+H

Query 434 CKQ--CGKAFNLSSSFYHERIH 454

32.94 (bits) f: 462 t: 482 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

Query *CpwPDGgKtFrrwsNLrRHMRTTH*

C+ CGK+FR++S+LR H TH

dkfzphtes3 462 CKQ--CGKAFRSASQLRVHGGTH 482

Query f: 490 t: 510 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

HMM *CpwPDGgKtFrrwsNLrRHMRTTH*

C++ CGK+FR+ S+LR H RTH

Query 490 CKE--CGKAFRSTSHLRVHGRTTH 510

30.69 (bits) f: 518 t: 540 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

Query *CpwPDGgKtFrrwsNLrRHMRTTH*

C++ CGK+FR+ +L++H R H

dkfzphtes3 518 CKE--CGKAFRYVKHLQIHERTE-KH 540

Query f: 552 t: 572 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

HMM *CpwPDGgKtFrrwsNLrRHMRTTH*

C++ C+K F ++ ++++H +TH

Query 552 CSI--CEKGFYSAKSFQTHEKTH 572

31.33 (bits) f: 580 t: 600 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

Query *CpwPDGgKtFrrwsNLrRHMRTTH*

C+ CGK+FR +LR H RTH

dkfzphtes3 580 CNQ--CGKAFRCCNSLRYHERTH 600

Query f: 608 t: 628 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

Alignment to HMM consensus:

HMM *CpwPDGgKtFrrwsNLrRHMRTTH*

C+ CGK+FR++S+LR+H RTH

Query 608 CKQ--CGKAFRSASHLRMHHERTH 628

35.30 (bits) f: 636 t: 656 Target: dkfzphtes3_10i1b.1
similarity to ZK1 (Homo sapiens), complete cds.

5

10

15

25

30

35

40

45

50

55

SEQ NPMNVRNAKKHSVNSLLFKYMKGSTEERSPMNVSIIVGMDSHLPRFFKYMQEHTLERNTMN
PRD ccccccccccccccccccccccccccccccccccchhhhhhhhhhhhhhcccc

SEQ VRNAEKHSIIFLPCIYTQGLIWERSHMNVRIVGKHSASLVPFMDMNRLTLEGSTMNASNV
PRD chhhhhhheeeccceeechhhhhccccceeeccccceeeccchhhhhhcccccccccc

5 SEQ AKLSHFVLFDIMKGLTLGRNPINVSSVGKPSFLLLLFNVMKGLTRERNPMSVF
PRD cccccccchhhhhhccccccccccccccccchhhhhhcccccccccccccc

(No Prosite data available for DKFZphtes3_10116.2)

10

(No Pfam data available for DKFZphtes3_10116.2)

DKFZphtes3_l0n10

5 group: testis derived

DKFZphtes3_l0n10 encodes a novel 502 amino acid protein without similarity to known proteins.

10 The mRNA is differentially polyadenylated and the novel protein is ubiquitously expressed.
No informative BLAST results; No predictive prosite, pfam or SCOP motive.

15 The new protein can find application in studying the expression profile of testis-specific genes.

unknown protein

20 differentially polyadenylated

Sequenced by Qiagen

25 Locus: unknown

Insert length: 2551 bp

Poly A stretch at pos. 2531, polyadenylation signal at pos. 2513

30

```

      1 CTCAGCCTCC CAAGTGGCTG GGACTGCAGG TTCTAAATGG CTTCTAAGAA
     51 GTTGGGTGCA GATTTTCATG GGACTTTCAG TTACCTTGAT GATGTCCCAT
    101 TTAAGACAGG AGACAAATTC AAAACACCAG CTAAAGTTGG TCTACCTATT
    151 GGCTTCTCCT TGCCTGATTG TTTGCAGGTT GTCAGAGAAG TACAGTATGA
    35 201 CTTCTCTTTG GAAAAGAAAA CCATTGAGTG GGCTGAAGAG ATTAAGAAAA
     251 TCGAAGAAGC CGAGCGGGAA GCAGAGTGCA AAATTGCGGA AGCAGAAGCT
    301 AAAGTGAATT CTAAGAGTGG CCCAGAGGGC GATAGCAAAA TGAGCTTCTC
    351 CAAGACTCAC AGTACAGCCA CAATGCCACC TCCTATTAAC CCCATCCTCG
    401 CCAGCTTGCA GCACAACAGC ATCCTCACAC CAACTCGGGT CAGCAGTAGT
    40 451 GCCACGAAAC AGAAAAGTTCT CAGCCCACCT CACATAAAGG CGGATTTCAA
     501 TCTTGCTGAC TTTGAGTGTG AAGAAGACCC ATTTGATAAT CTGGAGTTAA
     551 AAAGTATTGA TGAGAAGGAA GAGCTGAGAA ATATTCTGGT AGGAACCACT
    601 GGACCCATTA TGGCTCAGTT ATTGGACAAT AACTTGCCCA GGGGAGGCTC
    651 TGGGTCTGTG TTACAGGATG AGGAGGTCCT GGCATCCTTG GAACGGGCAA
    45 701 CCCTAGATTT CAAGCCTCTT CATAAACCCA ATGGCTTTAT AACCTTACCA
     751 CAGTTGGGCA ACTGTGAAAA GATGTCACTG TCTTCCAAAG TGTCCTCTCC
    801 CCCTATACCT GCAGTAAGCA ATATCAAATC CCTGTCTTTC CCCAAACTTG
    851 ACTCTGATGA CAGCAATCAG AAGACAGCCA AGCTGGCGAG CACTTTCAT
    901 AGCACATCCT GCCTCCGCAA TGGCACGTTT CAGAATTCCC TAAAGCCTTC
    50 951 CACCCAAAGC AGTGCCAGTG AGCTCAATGG GCATCACACT CTTGGGCTTT
   1001 CAGCTTTGAA CTTGGACAGT GGCACAGAGA TGCCAGCCCT GACATCCTCC
   1051 CAGATGCCTT CCCTCTCTGT TTTGTCTGTG TGCACAGAGG AATCATCACC
   1101 TCCAAATACT GGTCCCACGG TCACCCCTCC TAATTTCTCA GTGTCACAAG
   1151 TGCCCAACAT GCCCAGCTGT CCCCAGGCCT ATTCTGAACT GCAGATGCTG
    55 1201 TCCCCAGCG AGCGGCAGTG TGTGGAGACG GTGGTCAACA TGGGCTACTC
   1251 GTACGAGTGT GTCCTCAGAG CCATGAAGAA GAAAGGAGAG AATATTGAGC
   1301 AGATTCTCGA CTATCTCTTT GCACATGGAC AGCTTTGTGA GAAGGGCTTC
   1351 GACCCTCTTT TAGTGGAAGA GGCTCTGGAA ATGCACCAGT GTTCAGAAGA

```

1401 AAAGATGATG GAGTTTCTTC AGTTAATGAG CAAATTTAAG GAGATGGGCT
 1451 TTGAGCTGAA AGACATTAAG GAAGTTTTCG TATTACACAA CAATGACCAG
 1501 GACAATGCTT TGGGAAGACCT CATGGCTCGG GCAGGAGCCA GCTGAGACCA
 1551 GGCCCTGCCT AGGCCCTGCC GCAGAACCAC CATCCCTGGG AGGCCCTGCA
 5 1601 GAGCCCACCT GTGGGGAAAG AGAAGGGGCA GCTTCCGGAT TTTCTTTTGG
 1651 GGGTTAGAAG GTCAGGTGTG GAGACTGCTC GCCAGTCTCT GTGAGCCTAG
 1701 GCCCTGAGCT GGGGAGGTGG GGAAGATTCT GGCATGTGAG TGCCCCCAGA
 1751 ACTGTCCTGG CTCCTTCCGT ATTAACGCA TTTGCATTTT GAGAAGTGTC
 1801 CTTCCCACTT CAGCCCTCCG GAGAGACTAC CTTAGTCTTT CTGGGGTGTT
 10 1851 TATGTCCTCA GCTGAAGCCT GGCCTAGTTG CTGAGAGGGG CTGGGGAGAT
 1901 GGGGCGGGAG GGCCAGACTC AGTGCTGCTG TGGAGCTAGG TGCTTCCCCC
 1951 TTCCCTGAG ACTGGTTGAC TGAAGTCCAG TCAAGTTGAG TTCAAGTGAA
 2001 AGATTCTTCC AGGGTTTTAT TTTTCCCCCT CTAACAAAG TCTCATAGTG
 2051 TTAACACTGG TTCTGCAATA TCTCTGAGGT GCAAAGAATG CACTTTTCCC
 15 2101 TATGGGGCCC AGAGTTTGCC TTTTCTGCCA GGCAGTCACC ACGCTTCCCT
 2151 ACCCCAGCCT GTTTCTTTTG GCTTGTTTG GACCACAGTC CTCTGCTACC
 2201 CAGGGTTTTA GAGCCCTGCT TCTAGGAAAC AGTTTAAGAA ATCATTGGCC
 2251 CCTTCCCAGC ACATTGAATG GGTAAGCAGA CAGGCCATGA TTTAGTTGGC
 2301 CAGCACTAAT TCCACCTCTG TTCTCCTTGA ACAGCTTCCC CTCCAGCCCA
 20 2351 CTGCTTTAGG ATGACACAAT GAATAACACC TAGTCATAGA AATCAGTCTC
 2401 TCTGGTTTGT TTTGTATTAT GTTGACATC ATTAAAGATC TAAATACAAA
 2451 GGATATACAG TCTTGAATCT AAAATAATTT GCTAACTATT TTGATTCTTC
 2501 AGAGAGAACT ACTAATAAAA ATCTAAAAGG TAAAAAATAA AAAAAAATAA
 2551 A

BLAST Results

30 No BLAST result

Medline entries

35 No Medline entry

Peptide information for frame 1

ORF from 37 bp to 1542 bp; peptide length: 502

Category: putative protein

45 Classification: unclassified

1 MASKKLGADF HGTFSYLDVV PFKTGDKFKT PAKVGLPIGF SLPDCLQVVR
 51 EVQYDFSLEK KTIWAEIEIK KIEEAEREAE CKIAEAEAKV NSKSGPEGDS
 101 KMSFSKTHST ATMPPPINPI LASLQHNSIL TPTRVSSSAT KQKVLSPPHI
 50 151 KADFNLDFFE CEEDPFQNLK LKTIIDEKEEL RNILVGTTGP IMAQLLDNNL
 201 PRGGSGSVLQ DEEVLASLER ATLDLFKPLHK PNGFITLPQL GNCEKMSLSS
 251 KVSLLPIPAV SNIKSLSPFK LDSDDSNQKT AKLASTFHST SCLRNGTFQN
 301 SLKPSTQSSA SELNGHHTLG LSALNLDSTG EMPALTSSQM PSLSVLSVCT
 351 EESSPPNTGP TVTPPNFSVS QVPNMPSCPQ AYSELQMLSP SERQCVETVV
 55 401 NMGYSYECVL RAMKKKGNI EQILDYLFQAH GQLCEKGFDP LLVEEALEMH
 451 QCSEKMMEF LQLMSKFKEM GFELKDIKEV LLLHNNDDQN ALEDLMARAG
 501 AS

BLASTP hits

5 No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_10n10, frame 1

No Alert BLASTP hits found

Pedant information for DKFZphtes3_10n10, frame 1

Report for DKFZphtes3_10n10.1

```

[LENGTH]      502
[MMW]          55083.78
[pI]           5.02
[BLOCKS]      PR01083D
[BLOCKS]      BL01306B
[KW]           All_Alpha
[KW]           LOW COMPLEXITY      A-52 %

```

```
SEQ MASKKLGADFHGTFSYLDDVPFKTGDKFKTPAKVGLPIGFSLPDCLQVVREVQYDFSLEK
SEG .....xx
PRD cccccccccccccccccccccccccccccccccccccccchhhhhhhhhhccccch
```

```
SEQ KTIEWAAEEIKKIEEAEREAECKIAEAEAKVNSKSGPEGDSKMFSKTHSTATMPPPINPI  
SEG xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.....  
PRD hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccccccccccccccccccccccchhh
```

```
SEQ LASLQHNSILTPTRVSSSATKQKVLSPPHIKADFNLA DFECEEDPF DNLELKTIDEKEEL
SEG .....
PRD hhhhhccccccccccccccccchhhhhccccccchhhhhccccccccccccccccccccchhhhhhhh
```

```
SEQ  RNILVGTTPIMAQLLDNNLPRGGSGSVLQDEEVLASLERATLDFKPLHKPNGFITLPQL
SEG  .....
PRD  hhhhhccccchhhhhhhhhccccccccccccchhhhhhhhhhhhhhhcccccccccccccccccc
```

```
SEQ GNCEKMSLS$KVSLPPIPAVSNIKSLSFPKLDSDDSNQKTAKLASTFHSTSCLRNGTFQN
SEG .....
PRD CCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCCCCCCCCC
```

```
SEQ SLKPSTQSSASELNGHHTLGLSALNLD SGTEMPALTSSQMPSLSVLSVCTEESSPPNTGP
SEG .....xxxxxxx
PRD cccccccccccccccccccccccceccccccccccccccccccccccccccccccccccc
```

```
SEQ TVTPPNFSVSQVPNMPSCPAAYSELQMLSPSERQCVETVVNMGYSECVLRAMKKKGENI
SEG xxxxxx.....
PRD cccccccccccccccccccchhhhhhccccchhhhhhccccchhhhhhhhhhhcccbb
```

```
SEQ EQILDYLFAGHQLCEKGFDPLLVEEALEMHQCSSEKMMEFLQLMSKFKEMGFELKDIKEV
SEG .....
PRD hhhhhhhhhhhhccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
```

SEQ LLLHNNDDQDNALEDLMARAGAS

SEG
PRD hhccccchhhhhhhhhhhccc

5 (No Prosite data available for DKFZphtes3_10n10.1)

(No Pfam data available for DKFZphtes3_10n10.1)
DKFZphtes3_11a17

10

group: transmembrane protein

15 DKFZphtes3_11a17 encodes a novel 428 amino acid protein without
similarity to known proteins.

The novel protein contains 2 transmembrane regions and one
leucine zipper. The protein is ubiquitously expressed with higher
abundance in stomach, brain and testis.

20 No informative BLAST results; No predictive prosite, pfam or SCOP
motife.

The new protein can find application in studying the expression
profile of testis-specific genes and as a new marker for
25 testicular cells.

unknown protein

30 Pedant: TRANSMEMBRANE 2
perhaps differential polyadenylation

Sequenced by Qiagen

35 Locus: unknown.

Insert length: 2571 bp

Poly A stretch at pos. 2570, polyadenylation signal at pos. 2548

40

1 CTCTCCTGCG CCCTCTGGAG GAAGTGAGAA GAGTCAGTCC CACCCAGCTG
51 CCGCCTGGTA TCTGGGCTCC AGGCCACCGA GTATTTGGCC CCCAGCCACG
101 GAGCCCTTAG CACACACCTC CCCCACAGGT CCTGGAGATG TGGCTGAGCT
151 ACCTGCAGCC GTGGCGGTAC GCGCCTGACA AGCAGGCTCC GGGCAGCGAC
45 201 TCCCAAGCCC GGTGTGTGTC GGAGAAATGG GCACCCCTTG TCCAGGAGAA
251 CCTGCTGATG TACACCAAGT TGTGTGTGGG CTTTCTGAAC CGCGCGCTCC
301 GCACAGACCT GGTGAGCCCC AAGCACGCGC TCATGGTGTT CCGAGTGGCC
351 AAAGTCTTTG CCCAGCCCAA CCTGGCTGAG ATGATTCAGA AAGGTGAGCA
401 GCTATTCCTG GAGCCAGAGC TGGTCATCCC CCACCGCCAG CACCGACTCT
50 451 TCACGGCCCC CACATTCACT GGGAGCTTCC TGTCACCCTG GCCACCAGCG
501 GTCACTGATG CCTCCTTCAA GGTGAAGAGC CACGTCTACA GCCTGGAGGG
551 CCAGGACTGC AAGTACACCC CGATGTTTGG GCCCGAGGCC CGCACCTGG
601 TCCTGCGCCT CGCTCAGCTC ATCACACAGG CCAAACACAC AGCCAAGTCC
651 ATCTCCGACC AGTGTGCGGA GAGCCCGGCT GGCCACTCCT TCCTCTCATG
55 701 GCTGGGCTTT AGCTCCATGG ACACCAATGG CTCCTACACA GCCAACGACC
751 TGGACGAGAT GGGGCAAGAC AGTGTCCGGA AGACAGATGA ATACCTGGAG
801 AAGGCCCTGG AGTACCTGCG CCAGATATTC CGGCTCAGCG AAGCGCAGCT
851 CAGGCAGTTC AACTCGCCT TGGGCACCAC CCAGGATGAG AATGGAAAAA

901 AGCAACTCCC CGACTGCATC GTGGGTGAGG ACGGACTCAT CCTTACGCCC
951 CTGGGGCGGT ACCAGATCAT CAATGGGCTG CGAAGGTTTG AAATTGAGTA
1001 CCAGGGGGAC CCGGAGCTGC AGCCCATCCG GAGCTATGAG ATCGCCAGCT
1051 TGGTCCGCAC ACTCTTTAGG CTGTCGTCTG CCATCAACCA CAGATTTGCA
5 1101 GGACAGATGG CGGCTCTGTG TTCCCGGGAT GACTTCCTCG GCAGCTTCTG
1151 TCGCTACCAC CTCACAGAAC CTGGGCTGGC CAGCAGGCAC CTGCTGAGCC
1201 CTGTGGGGCG GAGGCAGGTG GCCGGCCACA CCCGCGGCCC CAGGCTCAGC
1251 CTGCGCTTCC TGGGCAGTTA CCGGACGCTG GTCTCGCTGC TGCTGGCCTT
1301 CTTCTGTGGC TCTCTGTTCT GCGTCGGGCC CCTCCCATGC ACGCTGCTGC
10 1351 TCACCCTGGG CTATGTCCTC TACGCCTCTG CCATGACACT GCTGACCGAG
1401 CGGGGGGAAGC TGCACCAGCC CTGAAGGTGT CAGCTGCCTT CAGAGCAGGC
1451 TGGAGGGATT TGCCACACAG CCCCACCCTT GGGCTGAGAG GACCTGGGAA
1501 GCCCCCTCAG GAGGGAACAC GGTCACTCTC GGGCTTCTGG AGCGGGGTTT
1551 CTGCAGCCGC AGAGGCATCT GGAGGAAACG CAACCAAGAA AGGAAGGCAG
15 1601 GTGGGCCCCA GCAAAGGAGT AGCTGCCAGG GCTCAACAGC TACGCTCTGT
1651 GACAGCGCAG AGCTCAGCGC CGGCCTTTCC CTCCCTCCGC CAAGGACTCA
1701 CGGCCAAGCC AGCTCTCGGG GCCTTTTTTC CAGTGCCCAT TTGGCTACTC
1751 TGCTGCACCA AGCTTGGGAG CCAGCCTGCC AACAGCCACC TGGGCCTGGC
1801 CTCCCCACTG GCTGGCCTTG AGGTTGGCAG AGTGGGTTGT GCGCCTTCCT
20 1851 CTCTCTGTGT GGGACCAGGA CAGTGGCTTA AGTCTCCACT CCAGGAAAGA
1901 ATCAAAGTTT CTAGAGTTGT GAGAAAACCA GAGAGTGGCT GTCTTGATTC
1951 TTCACTGTGA GGGGCGTTCT TCATGTTCTC CCAGCTGTTT CAAGACTGGG
2001 CCGTAGAATT CCATGTTTCA GGAGCCTAAG ACCCTCCAG AGCCCAGGGG
2051 CTTCAACGCA GACCCCAAGC CATTGAGCAC ATCACC AAA GCAGTGGCCA
25 2101 ACATCGCGGA CCCCTGTGCC TTGTCACAGA TGGGTGCTGG TCCTCAGGCG
2151 TTGGGGACAC TGCTGGGTCG ATGGGGTCGG ATTCTGCCAG TTTCTGCTCT
2201 GCAGCCAAAG ATGGTCAGAA GCATTGTCAC TTCAGTAACA TCAAGTGCTC
2251 AAAGACATGG CAACCGTTCA GTGGTACTTA AGTATTCAA ATATACA ACT
2301 ACAGATTCTC TGACAGAAAC CAGCACGGGG TCTTACCTT CATTACCCC
30 2351 ACAGGCGACA TGCAGGGGAG AACAGCATCT CAGTGGTGAT TTCCAAACCA
2401 AGCCTTTGTT TTCGGTGTGG GGTTTTGGGG GTTTGCTTTA ATGTTTTTGA
2451 AATTGTAAAT GTTGGGCTTT TTATTTTGAT GTAAACTGAG AATAATGGCA
2501 TTTTAGGGCC TGTGACCAA AATGAAGCTT GTAACGACCA TGGATCTGAA
2551 TAAACATGTC CTTGCTTCTG AAAAAAAAAA AAAAAAAAAA A

BLAST Results

40 Entry AF052134 from database EMBLNEW:
Homo sapiens clone 23585 mRNA sequence.
Score = 5765, P = 2.9e-254, identities = 1155/1156
3' UTR

45

Medline entries

50 No Medline entry

Peptide information for frame 3

55

ORF from 138 bp to 1421 bp; peptide length: 428
Category: putative protein

Classification: Transmembrane proteins unclassified
 Prosite motifs: LEUCINE_ZIPPER (404-425)

```

5      1 MWLSYLQPWRYAPDKQAPGSDSQPRCVSEK WAPFVQENLL MYTKLFVGF
51 NRALRTDLVS PKHALMVFRV AKVFAQPNLA EMIQKGEQLF LEPELVIPHR
101 QHRLFTAPT TGSFLSPWPP AVTDASFVKV SHVYSLEGQD CKYTPMFGPE
151 ARTLVRLAQ LITQAKHTAK SISDQCAESP AGHSFLSWLG FSSMDTNGSY
201 TANDLDEMGO DSVRKTDEYL EKALEYLRQI FRLSEAQLRQ FTLALGTTQD
10 251 ENGKKQLPDC IVGEDGLILT PLGRYQIING LRRFEIEYQG DPQLQPIRSY
301 EIASLVRTL FRLSSAINHRF AGQMAALCSR DDFLGSCRY HLTEPGLASR
351 HLLSPVGRRQ VAGHTRGPRL SLRFLGSYRT LVSLLLAFFV ASLFCVGPLP
401 CTLLLTG YV LYASAMTLLT ERGKLHQP
  
```

15

BLASTP hits

No BLASTP hits available

20

Alert BLASTP hits for DKFZphtes3_11a17, frame 3

No Alert BLASTP hits found

25

Pedant information for DKFZphtes3_11a17, frame 3

Report for DKFZphtes3_11a17.3

30

```

[LENGTH] 428
[MW]      48274.93
[pI]      8.92
[PROSITE] LEUCINE_ZIPPER 1
35 [KW]     TRANSMEMBRANE 2
    [KW]     LOW_COMPLEXITY 7.48 %
  
```

40

```

SEQ MWLSYLQPWRYAPDKQAPGSDSQPRCVSEK WAPFVQENLL MYTKLFVGF LNRA
SEG .....
PRD ccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
MEM .....
  
```

45

```

SEQ PKHALMVFRVAKVFAQPNLAEMIQKGEQLFLEPELVIPHRQHRLFTAPTFTGSFLSPWPP
SEG .....
PRD ccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
MEM .....
  
```

50

```

SEQ AVTDASFVKVSHVYSLEGQDCKYTPMFGPEARTLVRLRLAQ LITQAKHTAKSISDQCAESP
SEG .....
PRD ccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
MEM .....
  
```

55

```

SEQ AGHSFLSWLGFSSMDTNGSYTANDLDEMGO DSVRKTDEYLEKALEYLRQIFRLSEAQLRQ
SEG .....
PRD ccccccccccccccccccccccccccccccccccccccccccccccccccccccccc
MEM .....
  
```

Prosites for DKFZphtes3_11a17.3

(No Pfam data available for DKFZphtes3_11a17.3)

DKFZphtes3_11c22

5 group: signal transduction

DKFZphtes3_11c22 encodes a novel 482 amino acid protein with partial similarity to mouse PC32b.

10 The novel protein contains WD-repeats. WD-repeat proteins are known as regulatory elements in a large variety of pathways. The repeats form a propeller like structure, which serves as a platform for protein/protein interaction. The new protein is ubiquitously expressed, indicating that it takes an essential

15 regulatory function in the cell.

The new protein can find application in modulating/blocking of regulatory pathways.

20 similarity to mouse PC32b

perhaps complete cds.
contains WD-Repeats: cf. BLASTX-S37b94

25 perhaps differential polyadenylation

Sequenced by Qiagen

Locus: /map="1q23.2-24.3"

30 Insert length: 1952 bp
Poly A stretch at pos. 1932, polyadenylation signal at pos. 1912

35 1 GAAGCAAGTG AGGTTGCACA AAGCAATAGA GGACGAGGAA GATCTCGACC
51 CAGAGGTGGA ACAAGTCAAT CAGATATTTT AACTCTTCCT ACGGTCCCAT
101 CAAGTCCTGA TTTGGAAGTG AGTGAAACTG CAATGGAAGT AGATACTCCA
151 GCTGAACAAT TTCTTCAGCC TTCTACATCC TCTACAATGT CAGCTCAGGC
201 TCATTTCGACA TCATCTCCCA CAGAAAAGCCC TCATTCTACT CCTTTGCTAT
40 251 CTTCTCCAGA TAGTGAACAA AGGCAGTCTG TTGAGGCATC TGGACACCAC
301 ACACATCATC AGTCTGATTC TCCTTCTTCT GTGGTTAACA AACAGCTCGG
351 ATCCATGTCA CTTGACGAGC AACAGGATAA CAATAATGAA AAGCTGAGCC
401 CCAAACCAGG GACAGGTGAA CCAGTTTTAA GTTTGCACTA CAGCACAGAA
451 GGAACAAC TA CAAGCACAAT AAAACTGAAC TTTACAGATG AATGGAGCAG
45 501 TATAGCATCA AGTTCTAGAG GAATTGGGAG CCATTGCAAA TCTGAGGGTC
551 AGGAGGAATC TTTCTGTCCA CAGAGCTCAG TGCAACCACC AGAAGGAGAC
601 AGTGAAACAA AAGCTCCTGA AGAATCATCA GAGGATGTGA CAAAATATCA
651 GGAAGGAGTA TCTGCAGAAA ACCCAGTTGA GAACCATATC AATATAACAC
701 AATCAGATAA GTTCACAGCC AAGCCATTGG ATTCCAATC AGGAGAAAGA
50 751 AATGACCTCA ATCTTGATCG CTCTTGTTGG GTTCCAGAAG AATCTGCTTC
801 ATCTGAAAAA GCCAAGGAAC CAGAAACTTC AGATCAGACT AGCACTGAGA
851 GTGCTACCAA TGAAAATAAC ACCAATCCTG AGCCTCAGTT CCAAACAGAA
901 GCCACTGGGC CTTCAGCTCA TGAAGAAACA TCCACCAGGG ACTCTGCTCT
951 TCAGGACACA GATGACAGTG ATGATGACCC AGTCCTGATC CCAGGTGCAA
55 1001 GGTATCGAGC AGGACCTGGT GATAGACGCT CTGCTGTTGC CCGTATTAG
1051 GAGTTCTTCA GACGGAGAAA AGAAAGGAAA GAAATGGAAG AATTGGATAC
1101 TTTGAACATT AGAAGGCCGC TAGTAAAAAT GGTTTATAAA GGCCATCGCA
1151 ACTCCAGGAC AATGATAAAA GAAGCCAATT TCTGGGGTGC TAAC TTTGTA

```

1201 ATGAGTGGTT CTGACTGTGG CCACATTTTC ATCTGGGATC GGCACACTGC
1251 TGAGCATTTG ATGCTTCTGG AAGCTGATAA TCATGTGGTA AACTGCCTGC
1301 AGCCACATCC GTTTGACCCA ATTTTAGCCT CATCTGGCAT AGATTATGAC
1351 ATAAAGATCT GGTCAACATT AGAAGAGTCA AGGATTTTAA ACCGAAAAC
5 1401 TGCTGATGAA GTTATAACTC GAAACGAACT CATGCTGGAA GAAACTAGAA
1451 ACACCATTAC AGTTCAGCC TCTTTCATGT TGAGGATGTT GGCTTCACTT
1501 AATCATATCC GAGCTGACCG GTTGGAGGGT GACAGATCAG AAGGCTCTGG
1551 TCAAGAGAAT GAAAATGAGG ATGAGGAATA ATAAACTCTT TTTGGCAAGC
1601 ACTTAAATGT TCTGAAATTT GTATAAGACA TTTATTATAT TTTTTCTTT
10 1651 ACAGAGCTTT AGTGCAATTT TAAGGTTATG GTTTTGGAG TTTTCCCTT
1701 TTTTGGGAT AACCTAACAT TGGTTTGGAA TGATTGTGTG CATGAATTTG
1751 GGAGATTGTA TAAAACAAAA CTAGCAGAAT GTTTTAAAAA CTTTTGCCG
1801 TGTATGAGGA GTGCTAGAAA ATGCAAAAGT CAATATTTTC CCTAACCTTC
1851 AAATGTGGGA GCTTGGATCA ATGTTGAAGA ATAATTTTCA TCATAGTGAA
15 1901 AATGTTGGTT CAAATAAATT TCTACACTTG CCAAAAAAAA AAAAAAAA
1951 AA

```

BLAST Results

20

Entry HS702J19 from database EMBL:

Human DNA sequence *** SEQUENCING IN PROGRESS *** from clone 702J19

25 Score = 2043, P = 5.8e-252, identities = 425/445
10 exons matching Bp 316-1932

Entry HS536148 from database EMBL:

human STS WI-6347.

30 Score = 1203, P = 1.5e-47, identities = 247/252

Entry HS703H14 from database EMBLNEW:

Human DNA sequence from clone 703H14 on chromosome 1q23.2-24.3

35 Score = 1307, P = 1.1e-51, identities = 263/265
2 exons matching Bp 1-316

Medline entries

40

93026383:

Bergsagel PL, Timblin CR, Eckhardt L, Laskov R, Kuehl WM.;

Sequence and

45 expression of a murine cDNA encoding PC32b, a novel
gene expressed in plasmacytomas but not normal plasma cells.

Oncogene

1992 Oct;7(10):2059-64

50

Peptide information for frame 1

55

ORF from 133 bp to 1578 bp; peptide length: 482

Category: similarity to known protein

Classification: Protein management

Prositate motifs: MYB_1 (410-418)

```

      1 MEVDTPAEQF LQPSTSSSTMS AQAHSSTSSPT ESPHSTPLLS SPDSEQRQSV
5     51 EASGHHTHHQ SDSPSSVVNK QLGMSLDEQ QDNNNEKLSP KPGTGEPVLS
      101 LHYSTEGTTT STIKLNFTDE WSSIASSSRG IGSCHKSEGQ EESFVPQSSV
      151 QPPEGDSETK APEESSEDVT KYQEGVSAEN PVENHINITQ SDKFTAKPLD
      201 SNSGERNDLN LDRSCGVPEE SASSEKAKEP ETSDDTSTES ATNENNTNPE
      251 PQFQTEATGP SAHEETSTRD SALQDTDDSD DDPVLIPGAR YRAGPGDRRS
10    301 AVARIQEFFR RRKERKEMEE LDTLNIRRPL VKMVYKGHRN SRTMIKEANF
      351 WGANFVMSG SDCGHIFIWDR HTAEHMLLE ADNHHVNCLEQ PHPFDPILAS
      401 SGIDYDIKIW SPLEESRIFN RKLADDEVITR NELMLEETRN TITVPASFML
      451 RMLASLNHIR ADRLEGDRSE GSGQENENED EE

```

15

BLASTP hits

No BLASTP hits available

20

Alert BLASTP hits for DKFZphtes3_11c22, frame 1

TREMBLNEW:HS06631_1 gene: "H326"; Human (H326) mRNA, complete cds., N

25 = 1, Score = 278, P = 4e-22

PIR:S37694 gene PC326 protein - mouse, N = 1, Score = 265, P = 2.9e-20

30 PIR:T05676 hypothetical protein F20M13.40 - Arabidopsis thaliana, N =

1, Score = 240, P = 6.3e-18

35 >TREMBLNEW:HS06631_1 gene: "H326"; Human (H326) mRNA, complete cds.

Length = 597

HSPs:

40

Score = 278 (41.7 bits), Expect = 4.0e-22, P = 4.0e-22
Identities = 63/148 (42%), Positives = 94/148 (63%)

Query: 335 YKGRNSRTMIKEANFWG--

45 ANFVMSGSDCGHIFIWDRHTAEHMLLEADNH-VVNCLEP 391

YKGRN+ T +K NF+G + FV+SGSDCGHIF+W++ + + + +E D

VVNCLEP

Sbjct: 428 YKGRNNAT-

VKGVNFYGPKEFVVSQSDCGHIFLWEKSSCQIIQFMEGDKGGVVNCLEP 486

50

Query: 392 HPFDPILASSGIDYDIKIWSPLEESRIFNRKLADDEVITRNEMLLEE-
TRNTITVPASFML 450

HP P+LA+SG+D+D+KIW+P E+ L D VI +N+ +E + +

+ S ML

55 Sbjct: 487 HPHLPVLATSGLDHDKIWAAPTAEASTELTGLKD-

VIKKNKRERDEDSLHQTDLFDSHML 545

Query: 451 RMLASLNHIRADRLEGD-RSESGQENENEDE 481

L ++H+R R R G G + + DE

Pedant information for DKFZphtes3_11c22, frame 1

```

[LENGTH] 482
[MW] 53470.92
[pI] 4.72
[HOMOL] PIR:T04961 hypothetical protein T12J5.10 -
Arabidopsis thaliana 2e-22
[FUNCAT] 30.09 organization of intracellular transport vesicles
[S. cerevisiae, YDL145c] 4e-05
[FUNCAT] 08.07 vesicular transport (golgi network, etc.) [S.
cerevisiae, YDL145c] 4e-05
[FUNCAT] 99 unclassified proteins [S. cerevisiae, YCLO39w]
2e-04
[SUPFAM] WD repeat homology 4e-21
[PROSITE] MYB_1 1
[KW] Alpha_Beta
[KW] LOW_COMPLEXITY 17.01 %

```

[illegible]

SEQ EE
SEG ..
PRD cc

5

Prosites for DKFZphtes3_11c22.1

10 PS00037 410->419 MYB_1 PD0C00037

(No Pfam data available for DKFZphtes3_11c22.1)

DKFZphtes3_11d21

5 group: signal transduction

DKFZphtes3_11d21 encodes a novel 922 acid protein and contains the full coding sequence of the human Nedd-4-like ubiquitin-protein ligase.

10 The novel protein contains four WW domains. The WW/rsp5/WWP domain has been shown to bind proteins with particular proline-motifs, and thus resembles somewhat SH3 domains. It is frequently associated with other domains typical for proteins in
15 signal transduction processes. There is also a ubiquitin-protein ligase activity reported. The protein is believed to play an important role in protein-degradation pathways.

20 The new protein can find application in diagnosis of diseases due to unnormal protein degradation like muscular dystrophy or multiple sclerosis as well as in modulating the half life of specific proteins and in expression profiling.

25 similarity to Nedd-4-like ubiquitin-protein ligase (Homo sapiens)

Sequenced by Qiagen

Locus: unknown

30 Insert length: 3382 bp

Poly A stretch at pos. 3362, polyadenylation signal at pos. 3345

```

35      1 ATTTTGGGAC ATGGCCACTG CTTACCAAG GTCTGATACT AGTAATAACC
      51 ACAGTGGAAG GTTGCACTTA CAGGTAAGT TTTCTAGTGC CAAACTTAAA
     101 AGAAAAAAGA ACTGGTTTCGG AACAGCAATA TATACAGAAG TAGTTGTAGA
     151 TGGAGAAATT ACGAAAACAG CAAAATCCAG TAGTTCTTCT AATCCAAAT
     201 GGGATGAACA GCTAACTGTA AATGTTACGC CACAGACTAC ATTGGAATTT
     40  251 CAAGTTTGGA GCCATCGCAC TTTAAAAGCA GATGCTTTAT TAGGAAAAGC
     301 AACGATAGAT TTGAAACAAG CTCTGTTGAT ACACAATAGA AAATTGGAAA
     351 GAGTGAAAGA ACAATTAATA CTTTCCTTGG AAAACAAGAA TGGCATAGCA
     401 CAAACTGGTG AATTGACAGT TGTGCTTGAT GGATTGGTGA TTGAGCAAGA
     451 AAATATAACA AACTGCAGCT CATCTCCAAC CATAGAAATA CAGGAAAATG
     45  501 GTGATGCCTT ACATGAAAAT GGAGAGCCTT CAGCAAGGAC AACTGCCAGG
     551 TTGGCTGTTG AAGGCACGAA TGGGAATAGT AATCATGTAC CTACAAGCAC
     601 TCTAGTCCAA AACTCATGCT GCTCGTATGT AGTTAATGGA GACAACACAC
     651 CTTTCATCTCC GTCTCAGGTT GCTGCCAGAC CCAAAAATAC ACCAGCTCCA
     701 AAACCACTCG CATCTGAGCC TGCCGATGAC ACTGTTAATG GAGAATCATC
     50  751 CTCATTTGCA CCAACTGATA ATGCGTCTGT CACGGGTACT CCAGTAGTGT
     801 CTGAAGAAAA TGCCTTGTCT CCAAATTGCA CTAGTACTAC TGTGAAGAT
     851 CCTCCAGTTC AAGAAATACT GACTTCCTCA GAAAACAATG AATGTATTCC
     901 TTCTACCACT GCAGAATTGG AATCTGAAGC TAGAAGTATA TTAGAGCCTG
     951 ACACCTCTAA TTCTAGAAGT AGTTCCTGCT TTGAAGCAGC CAAATCAAGA
     55 1001 CAGCCAGATG GGTGTATGGA TCCTGTACGG CAGCAGTCTG GGAATGCCAA
     1051 CACAGAAACC TTGCCATCAG GGTGGGAACA AAGAAAAGAT CCTCATGGTA
     1101 GAACCTATTA TGTGGATCAT AATACTCGAA CTACCACATG GGAGAGACCA
     1151 CAACCTTTAC CTCCAGGTTG GGAAAGAAGA GTTGATGATC GTAGAAGAGT

```

1201 TTATTATGTG GATCATAACA CCAGAACAAC AACGTGGCAG CGGCCTACCA
1251 TGGAATCTGT CCGAAATTTT GAACAGTGGC AATCTCAGCG GAACCAATTG
1301 CAGGGAGCTA TGCAACAGTT TAACCAACGA TACCTCTATT CGGCTTCAAT
1351 GTTAGCTGCA GAAAATGACC CTTATGGACC TTTGCCACCA GGCTGGGAAA
5 1401 AAAGAGTGGG TTCAACAGAC AGGGTTTACT TTGTGAATCA TAACACAAAA
1451 ACAACCCAGT GGGAAAGATCC AAGAACTCAA GGCTTACAGA ATGAAGAACC
1501 CCTGCCAGAA GGCTGGGAAA TTAGATATAC TCGTGAAGGT GTAAGGTACT
1551 TTGTTGATCA TAACACAAGA ACAACAACAT TCAAAGATCC TCGCAATGGG
1601 AAGTCATCTG TAACTAAAGG TGGTCCACAA ATTGCTTATG AACGCGGCTT
10 1651 TAGGTGGAAG CTTGCTCACT TCCGTTATTT GTGCCAGTCT AATGCACTAC
1701 CTAGTCATGT AAAGATCAAT GTGTCCCGGC AGACATTGTT TGAAGATTCC
1751 TTCCAACAGA TTATGGCATT AAAACCCCTAT GACTTGAGGA GGCGCTTATA
1801 TGTAATATTT AGAGGAGAAG AAGGACTTGA TTATGGTGGC CTAGCGAGAG
1851 AATGGTTTTT CTTGCTTTCA CATGAAGTTT TGAACCCAAT GTATTGGTTA
15 1901 TTTGAGTATG CGGGCAAGAA CAACTATTGT CTGCAGATAA ATCCAGCATC
1951 AACCATTAAT CCAGACCATC TTTCATACTT CTGTTTCATT GGTCGTTTTA
2001 TTGCCATGGC ACTATTTTCT GGAAGTTT TCGATACTGG TTTCTCTTTA
2051 CCATTCTACA AGCGTATGTT AAGTAAAAAA CTTACTATTA AGGATTTGGA
2101 ATCTATTGAT ACTGAATTTT ATAACCTCCT TATCTGGATA AGAGATAACA
20 2151 ACATTGAAGA ATGTGGCTTA GAAATGTACT TTTCTGTTGA CATGGAGATT
2201 TTGGGAAAAG TTACTTCACA TGACCTGAAG TTGGGAGGTT CCAATATTCT
2251 GGTGACTGAG GAGAACAAAG ATGAATATAT TGGTTTAATG ACAGAATGGC
2301 GTTTTTCTCG AGGAGTACAA GAACAGACCA AAGCTTTCCT TGATGGTTTT
2351 AATGAAGTTG TTCCTCTTCA GTGGCTACAG TACTTCGATG AAAAAGAATT
25 2401 AGAGGTTATG TTGTGTGGCA TGCAGGAGGT TGAATTGGCA GATTGGCAGA
2451 GAAATACTGT TTATCGACAT TATACAAGAA ACAGCAAGCA AATCATTTGG
2501 TTTTGGCAGT TTGTGAAAGA GACAGACAAT GAAGTAAGAA TGCGACTATT
2551 GCAGTTCGTC ACTGGAACCT GCCGTTTACC TCTAGGAGGA TTTGCTGAGC
2601 TCATGGGAAG TAATGGGCCT CAAAAGTTTT GCATTGAAAA AGTTGGCAAA
30 2651 GACACTTGGT TACCAAGAAG CCATACATGT TTTAATCGCT TGGATCTACC
2701 ACCATATAAG AGTTATGAAC AACTAAAGGA AAAACTTCTT TTTGCAATAG
2751 AAGAGACAGA GGGATTTGGA CAAGAATGAA TGTGGCTTCT TATTTTGGAG
2801 GAGCTCTTGC ATTTAAATAC CCCAGCCAAG AAAAATTGCA CAGATAGTGT
2851 ATATAAGCTG TTCATTCTGT ACAGTGAATT TTCCGAACCT CTCAAAGTAT
35 2901 GTTTTCCGTT CTTCCACAGA AATATGCAAA ACAGTTCATC CTTTTCTACT
2951 TTATTTATTG TTCCCTTGAA ATGACTGACC AGGAAAAAGA TCATCCTTAA
3001 ATTTTGAAGC AAGTGAGAGA CTTTATTAAA AATACATATA TATCTATATA
3051 AACATATAAG ATAGTGGCTC TAGTTTTATA GAGCTCCAAG TGTATTAAAC
3101 ATGACAGCCA TTCATTCTA AAGATCTGGA TTTGCTTTAC CTTGTTAATA
40 3151 TTATCTAGGG GAAAAAGTGC AAATTGCTCC ATGTTCTTCT CTCCCTTATG
3201 TAACATCTCC TGAGGGTGTG TAGTTGCATG GCTGTTTACA AAGGTATTAA
3251 GGGCTTAGGC CAAATCTTAC TTTGAGTATG TTAACAAAAA AAAAATGCTG
3301 CTGGCTTTTC TGAAGACAGG TGCTTGAAC TGTGAGTTTG TTTTAAATAA
3351 ATACAATAGT TGAACAAAAA AAAAAAAA AA

45

BLAST Results

50 No BLAST result

Medline entries

55

97313427:

Pirozzi G, McConnell SJ, Uveges AJ, Carter JM, Sparks AB, Kay BK,
Fowlkes DM.; Identification of novel human WW domain-containing

proteins
by cloning of ligand targets. J Biol Chem 1997 Jun
6;272(23):14611-6

5

Peptide information for frame 2

10

ORF from 11 bp to 2776 bp; peptide length: 922

Category: known protein

Classification: Protein management

Prosite motifs: WW_DOMAIN_1 (355-380)

15

WW_DOMAIN_1 (387-412)

WW_DOMAIN_1 (462-487)

WW_DOMAIN_1 (502-527)

20

1 MATASPRSDT SNNHSGRLQL QVTVSSAKLK RKKNWFGTAI YTEVVVDGEI

51 TKTAKSSSSS NPKWDEQLTV NVTPQTTFLEF QVWSHRTLKA DALLGKATID

101 LKQALLIHNR KLERVKEQLK LSLENKNGIA QTGELTVVLD GLVIEQENIT

151 NCSSSPTIEI QENGDAIHEN GEPSARTTAR LAVEGTNGID NHVPTSTLVQ

201 NSCCSYVVNG DNTSPSSPSQV AARPKNTPAP KPLASEPADT TVNGESSSFA

25

251 PTDNASVTGT PVVSEENALS PNCTSTTVED PPVQEILTSS ENNECIPSTS

301 AELESEARSI LEPDTSNSRS SSFAEAAKSR QPDGCMOPVR QQSGNANTET

351 LPSGWEQRKD PHGRTYYVDH NTRTTTWERP QPLPPGWERR VDDRRRVYYV

401 DHNTRTTTWQ RPTMESVRNF EQWQSQRNQL QGAMQQFNQR YLYSASMLAA

451 ENDPYGPPLP GWKRVNSTD RYFVNHNTK TTQWEDPRTQ GLQNEEPLPE

30

501 GWEIRYTREG VRYFVDHNTR TTTFKDPRNG KSSVTKGGPQ IAYERGFRWK

551 LAHFRYLCQS NALPSHVKIN VSRQTLFEDS FQQIMALKPY DLRRRLYVIF

601 RGEGLDYGG LAREWFFLLS HEVLNPMYCL FEYAGKNNYC LQINPASTIN

651 PDHLSYFCFI GRFIAMALFH GKFI DTGFSL PFYKRMLSCK LTIKDLESID

701 TEFYNSLIWI RDNNIEECGL EMYFSVDMET LGKVTSHDLK LGGSNILLVTE

35

751 ENKDEYIGLM TEWRFSRGVQ EQTKAFLDGF NEVVPLQWLQ YFDEKELEVW

801 LCGMQEVDLA DWQRNTVYRH YTRNSKQIIW FWQFVKETDN EVRMRLQLQFV

851 TGTCRLPLGG FAELMGSNGP QKFCIEKVGK DTWLPRSHTC FNRLDLPPYK

901 SYEQLKEKLL FAIEETEGFG QE

40

BLASTP hits

No BLASTP hits available

45

Alert BLASTP hits for DKFZphtes3_11d21, frame 2

No Alert BLASTP hits found

50

Pedant information for DKFZphtes3_11d21, frame 2

Report for DKFZphtes3_11d21.2

55

[LENGTH] 925
[MW] 105650.58
[pI] 5.60

[HOMOL] TREMBL:HSU96113_1 product: "WWP1"; Homo sapiens
 Nedd-4-like ubiquitin-protein ligase WWP1 mRNA, partial cds. 0.0
 [FUNCAT] 30.02 organization of plasma membrane [S. cerevisiae,
 YER125w] 1e-149
 5 [FUNCAT] 11.01 stress response [S. cerevisiae, YER125w] 1e-
 149
 [FUNCAT] 06.13.01 cytoplasmic degradation [S. cerevisiae,
 YER125w] 1e-149
 [FUNCAT] 03.10 sporulation and germination [S. cerevisiae,
 10 YER125w] 1e-149
 [FUNCAT] 06.07 protein modification (glycosylation, acylation,
 myristylation, palmitylation, farnesylation and processing)
 [S. cerevisiae, YER125w] 1e-149
 [FUNCAT] 03.22 cell cycle control and mitosis [S. cerevisiae,
 15 YDR457w] 1e-78
 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YJR036c]
 7e-39
 [FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae,
 YKL010c] 8e-21
 20 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YKL012w]
 6e-05
 [FUNCAT] 04.05.03 mrna processing (splicing) [S. cerevisiae,
 YKL012w] 6e-05
 [FUNCAT] 30.01 organization of cell wall [S. cerevisiae,
 25 YIRO19c] 3e-04
 [FUNCAT] 30.90 extracellular/secretion proteins [S. cerevisiae,
 YIRO19c] 3e-04
 [FUNCAT] 01.05.01 carbohydrate utilization [S. cerevisiae,
 YIRO19c] 3e-04
 30 [BLOCKS] BP03746E
 [BLOCKS] BP03761G
 [BLOCKS] BL00514E Fibrinogen beta and gamma chains C-terminal
 domain proteins
 [BLOCKS] PR00731B
 35 [BLOCKS] BP01566C
 [BLOCKS] BL01159 WW/rsp5/WWP domain proteins
 [BLOCKS] PR00403B
 [BLOCKS] PR00403A
 [BLOCKS] PF00632B
 40 [BLOCKS] PF00632A
 [EC] 6.3.2.19 Ubiquitin--protein ligase 1e-151
 [PIRKW] ligase 1e-151
 [PIRKW] transmembrane protein 2e-37
 [PIRKW] leucine zipper 2e-28
 45 [SUPFAM] WW repeat homology 1e-151
 [SUPFAM] WD repeat homology 2e-28
 [SUPFAM] ubiquitin ligase homolog 1e-151
 [PROSITE] WW_DOMAIN_1 4
 [PFAM] WW/rsp5/WWP domain containing proteins
 50 [PFAM] C2 domain
 [KW] Alpha_Beta
 [KW] LOW_COMPLEXITY 1.41 %

55 SEQ FWDMATASPRSDTSNNHSGRLQLQVTVSSAKLKRKKNWFGTAIYTEVVVDGEITKTAKSS
 SEG
 PRD cccccccccccccccccceeeeeehhhhhhhhhhhcccccccccccccccccc

SEQ SSSNPKWDEQLTVNVTPQTTLFQVWSHRTLKADALLGKATIDLKQALLIHNRLERVKE
SEG
PRD ccc

5 SEQ QLKLSLENKNGIAQTGELTVVLDGLVIEQENITNCSSSPTIEIQENGDALHENGEPART
SEG
PRD hhhhhhcc

10 SEQ TARLAVEGTNGIDNHVPTSTLVQNSCCSYVVNGDNTSPSPSQVAARPKNTPAPKPLASEP
SEG
PRD hhhhhhcc

15 SEQ ADDTVNGESSFAPTNDASVTGTPVVSEENALSPNCTSTTVEDPPVQEIILTSSENNECIP
SEG
PRD ccc

20 SEQ STSAELESEARSILEPDTNSNRSSSAFEAAKSRQPDGCM DPVRQQSGNANTETLP SGWEQ
SEG
PRD ccc

25 SEQ RKDPHGRTYYVDHNTRTTTWERPQPLPPGWERRVDDRRRVYYVDHNTRTTTWQRPTMESV
SEG
PRD ccc

30 SEQ RNFEQWQSQRNQLQGAMQAFNORYLYSASMLAENDPYGPLPPGW EKRVDSTDRVYFVNH
SEG
PRD hhh

35 SEQ NTKTTQWEDPRTQGLQNEEPLPEGWEIRYTRGVRYFVDHNTRTTTFKDP RN GKSSVTKG
SEG
PRD ccc

40 SEQ GPQIAYERGFRWKL AHFRYLCQSNALPSHV KINVS RQTLFEDSFQIMALKPYDLRRRLY
SEG
PRD cccccchhh

45 SEQ VIFRGEGLDYGGLAREWFFLLSHEVLNPMYCLFEYAGKN NYCLQINPASTINPDHLSYF
SEG
PRD hhhccccccccccccchhh

50 SEQ CFIGRFIAMALFHGKFIDTGFSLPFYKRMLS KKLTIKDL ESIDTEFYNSLIWIRDNNIEE
SEG
PRD hhh

55 SEQ CGLEMYFSVDMEILGKV TSHDLKLGGSNILVTEENKDEYIGLMTEWRFSRGVQEQTKAFL
SEG
PRD chhh

SEQ DGFNEVVPLQWLQYFDEKELEVMLCGMQEVDLADWQRNTVYRHYTRNSKQIIWFQFVKE
SEG
PRD hhhhhccccchhh

SEQ TDNEVRMRLQFVTGTCRLPLGGFAELMG SNGPQKFCIEKVGKDTWLPRSHTCFNRLDLP
SEG
PRD hchhh

SEQ PYKSYEQ LKEKLLFAIEETEGFGQE
SEG

5 Prosite for DKFZphtes3_11d21.2

```

      PS01159      358->384      WW_DOMAIN_1      PD0C50020
      PS01159      390->416      WW_DOMAIN_1      PD0C50020
      PS01159      465->491      WW_DOMAIN_1      PD0C50020
10    PS01159      505->531      WW_DOMAIN_1      PD0C50020

```

Pfam for DKFZphtes3_11d21.2

HMM_NAME C2 domain

20 *L t V r I I e A R N L W k M D M n G f S D P Y V K V d M d P d p k D t k K W K T k T i W N N . G L
L V++ +A+ +K++++G+ Y +V +D+++ TKT

```
Query           23  LQVTVSSAKLKRKKNWFGTA-IYTEVVVDGE-----
ITKTAKSSSSSS   b3
```

HMM NPVWNEEfVfFedIPyPdIqrkMLRFaVWDWDRFSRBDFIGHCi*
NP W+ E+++ + + + L+F+VW + ++ + ++G ++
Query 64 NPKWD-EQLTVN---VTPQTT--LEFQVWSHRTLKADALLGKAT
100

HMM_NAME WW/rsp5/WWP domain containing proteins

```

35  HMM                      *LPSGWEeHWDpsGRpWYYWNHETkTTQWEP*
                                LPSGWE+++DP GR+ YY++H+T+TT+WE+P
Query          354  LPSGWEQRKDPHGRT-YYVDHNTRTTTWERP          383

```

50.09 386 415 1 31 dkfzphes3_11d21.2 similarity to
40 Nedd-4-like ubiquitin-protein ligase (Homo sapiens)
Alignment to HMM consensus:

Query		*LPsGWEeHWDpsGRpWYYWNHETkTTQWEP*	
		LP+GWE++ D+ R YY++H+T+TT+W++P	
dkfzphes3	386	LPPGWERRVDDRRRV-YYVDHNTRTTTWQRP	415

Query 490 1 31 dkfzphtes3_11d21.2 similarity to
Nedd-4-like ubiquitin-protein ligase (Homo sapiens)
Alignment to HMM consensus:

```

HMM      *LPsGWEeHWDpsGRpWYYWNHETkTTQWEpP*
          LP+GWE++ D + R Y++NH+TKTTQWE+P
Query    461  LPPGWEKRVDSTDRV-YFVNHNHTKTTQWEDP

```

38.62 501 530 1 31 dkfzphtes3_11d21.2 similarity to
Nedd-4-like ubiquitin-protein ligase (Homo sapiens)
Alignment to HMM consensus:

```

Query          *LPsGWEeHWDpsGRpWYYWNHETkTTQWEP*
               LP GWE +++ +G + Y+++H+T+TT+ ++P
dkfzphtes3    501  LPEGWEIRYTREGVR-YFVDHNTRTTTFKDP      530

```

PAGE INTENTIONALLY LEFT BLANK

5 group: testis derived

DKFZphtes3_11e17 encodes a novel 573 amino acid protein without similarity to known proteins.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motife.

The new protein can find application in studying the expression profile of testis-specific genes.

15

unknown protein

Sequenced by Qiagen

20

Locus: unknown

Insert length: 2102 bp

Poly A stretch at pos. 2080, polyadenylation signal at pos. 2059

25

```
1  GGCCTGGGGG GCTTCCCTGG GGGGCTTGTC GCCGGGGCCG CCTGGGCTTT
5  CAGGTCTTCC GAGGCTGACA TTCACGTTTC ATTCTGCCAC ACTCGGGAAC
10  GGTGATCGGG GAAGCATGGG GATCCGGGAG AAGCACCAC AAAACTAGCA
30 151 TCCTCCTGGA GGAGCTCGGG AATAGGATGA GTGATAATCC ACCCAGAATG
201 GAAGTGTGTC CTTACTGTAA GAAGCCATTT AAACGATTAA AATCCCACTT
251 GCCATACTGT AAGATGATAG GATCAACCAT ACCTACTGAT CAAAAAGTTT
301 ATCAGTCCAA GCCAGCTACA CTCCCACGTG CTAAAAAGAT GAAAGGACCA
351 ATCAAAGATT TAATTAAAGC TAAAGGGAAA GAGTTAGAGA CAGAGAATGA
35 401 AGAAAGAAAT TCTAAGTTGG TGGTGGACAA ACCAGAACAG ACAGTGAAGA
451 CCTTTCCACT GCCAGCTGTT GGTTTGGAAA GAGCAGCTAC TACAAAGGCA
501 GATAAAGACA TCAAGAATCC AATCCAACCA TCCTTCAAAA TGTTAAAAAA
551 TACTAAACCA ATGACTACTT TCCAAGAAGA AACCAAGGCT CAGTTTTACG
601 CATCAGAGAA AACCTCTCCT AAAAGAGAAC TTGCCAAAGA TTTGCCTAAA
40 651 TCAGGAGAAA GTCGATGTAA TCCTTCAGAA GCTGGAGCGT CTTTACTGGT
701 TGGCTCAATA GAACCTTCTT TGTCAAATCA AGATAGAAAA TATTCCTCAA
751 CTCTACCTAA TGATGTACAA ACTACCTCTG GTGATCTCAA ATTTGGACAA
801 ATTGATCCCC AAAGACAGGA ACTTCTAGTA AAATTACTAG ATGTGCCTAC
851 TGGTGATTGT CATATTTCTC CAAAGAATGT CAGTGATGGG GTTAAAAGGG
45 901 TAAGAACATT ATTAAGCAAT GAGAGAGATT CCAAAGGCAG GGATCACCTC
951 TCAGGAGTCC CTACTGATGT TACAGTTACT GAGACTCCAG AAAAGAACAC
1001 AGAATCCCTC ATTTTAAGCC TTAATATGAG CTCATTAGGT AAAATCCAAG
1051 TCATGGAGAA ACAAGAGAAA GGACTTACCC TGGGAGTAGA GACGTGTGGG
1101 AGCAAAGGAA ATGCAGAGAA AAGTATGTCT GCAACAGAAA AGCAGGAACG
50 1151 GACTGTCATG AGCCATGGCT GTGAGAACTT CAACACCAGG GATTCAGTCA
1201 CAGGAAAGGA GTCTCAAGGG GAAAGACCAC ATTTAAGTTT GTTCATTCCG
1251 AGGGAGACGA CTTACCAGTT TCATTCTGTA TCGCAGTCAA GTAGTCAAAG
1301 TCTTGCCCTC CTAGCTACAA CATTTCTTCA AGAAAAGAAA GCAGAAGCCC
1351 AGAATCATAA TTGTGTCCCT GATGTAAAGG CATTAATGGA GAGTCCCGAG
55 1401 GGACAGTTAT CTCTGGAGCC CAATCTGAT AGTCAGTTCC AAGCATCACA
1451 CACTGGGTGC CAGAGCCCTT TATGTTCAGC CCAGCGTCAC ATCCTCAGA
1501 GCCCCTTCAC CAATCATGCT GCAGCTGCTG GCAGGAAGAC TCTTCGCAGC
1551 TGCATGGGGC TGGAGTGGTT TCCAGAGCTC TATCCTGGTT ACCTTGGACT
```


1601 AGGGGTGTTG CCAGGGAAGC CTCAGTGTG GAATGCAATG ACCCAGAAGC
1651 CACAACCTTAT CAGTCCCCAG GGGGAAAGAC TCTCACAAGG CTGGATCAGG
1701 TGCAACACCA CCATAAGGAA GAGTGGATTG GGTGGCATCA CTATGCTCTT
1751 CACAGGATAC TTCGTCCTGT GTTGTAGCTG GAGTTTCAGA CGTCTGAAAA
5 1801 AATTGTGCCG ACCCCTGCCC TGGGAAGAGCA CAGTACCTCC ATGCATTGGT
1851 GTGGCGAAGA CGACTGGGGA TTGCCGCTCT AAAACATGTT TGGATTAGGA
1901 AGCACGTTTA AGTAGGAGAA GCCTTCGTGA CTTCTCTCTA GTGCCTTCGT
1951 GCCCTGTGTT GCCCACTGAA TTGCCCTGTA ACACCTAAGT GTAGTGGTAG
2001 CATTAAGGGA TAGCTTTTCA GCCCTCAAGG TTATCAGGAG CATTGTATC
10 2051 ACTGCTATAA ATAAAGTAGT ATCACTTGTC ATAAAAA AAAAAA
2101 AA

BLAST Results

15

No BLAST result

20

Medline entries

No Medline entry

25

Peptide information for frame 3

30 ORF from 177 bp to 1895 bp; peptide length: 573
Category: putative protein
Classification: no clue

35 1 MSDNPPRMEV CPYCKKPFKR LKSHLPYCKM IGSTIPTDQK VYQSKPATLP
51 RAKKMKGPIK DLIKAKGKEL ETENEERNSK LVVDKPEQTV KTFPLPAVGL
101 ERAATTKADK DIKNPIQPSF KMLKNTKPM TFEETKAQF YASEKTSPIR
151 ELAKDLPSKG ESRCNPSEAG ASLLVGSIEP SLSNQDRKYS STLPNDVQTT
201 SGDLKLDKID PQRQELLVKL LDVPTGDCHI SPKNVSDGVK RVRTLLSNER
251 DSKGRDHLSG VPTDVTVTET PEKNTESLIL SLKMSSLGKI QVMEKQEKGL
40 301 TLGVETCGSK GNAEKSMSAT EKQERTVMSH GCENFNTRDS VTGKESQGER
351 PHLSLFIPRE TTYQFHSVSQ SSSQSLASLA TTFLQEKKAQ AQNHNCVPDV
401 KALMESPEGQ LSLEPKSDSQ FQASHTGCQS PLCSAQHRTP QSPFTNHAAA
451 AGRKTLRSCM GLEWFPELYP GYLGLGVLPK KPQCWNAMTQ KPQLISPQGE
501 RLSQGWIRCN TTIRKSGFGG ITMLFTGYFV LCCSWSFRL KKLCRPLPWK
45 551 STVPPCIGVA KTTGDCRSKT CLD

BLASTP hits

50

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_11e17, frame 3

55 No Alert BLASTP hits found

Pedant information for DKFZphtes3_11e17, frame 3

Report for DKFZphtes3_11e17.3

5 [LENGTH] 573
 [MW] 63389.88
 [pI] 9.24
 [BLOCKS] BL00028 Zinc finger, C2H2 type, domain proteins
 [KW] Alpha_Beta
10 [KW] LOW_COMPLEXITY 7.50 %

 SEQ MSDNPPRMEVCPYCKKPFKRLKSHLPYCKMIGSTIPTDQKVYQSKPATLPRAKKMKGPIK
 SEG
15 PRD ccccccceeeccccchhhhhhhccceeeccccccccceeeccccchhhhhhhccch

 SEQ DLIKAKGKELETENEERNNSKLVVDKPEQTVKTFPLPAVGLERAATTKADKDIKNPIQPSF
 SEG
20 PRD hhhhhccccchhhhhhhheeeccccccccceccccchhhhhhhhhhhccccccccchh

 SEQ KMLKNTKPMTTFQETKAQFYASEKTSKRELAKDLPKSGESRCNPSEAGASLLVGSIEP
 SEG
 PRD hhhhccccchhhhhhhhhhhhhccccchhhhhccccccccccccccccchhhhhhhcccc

25 SEQ SLSNQDRKYSSTLPNDVQTTSGDLKLDKIDPQRQELLVKLLDVPTGDCHISPKNVSDGVK
 SEG
 PRD ccccccceeeccccccccccccccccccccccccchhhhhhhhhccccccccccccccccchh

 SEQ RVRTLSSNERDSKGRDHLSGVPTDVTVTETPEKNTESLILSLKMSSLGKIQVMEKQEKGL
30 SEG xxxxxxxxxxxx.....
 PRD hhhhhccccccccccccccccccccceeeccccchhhhhhhhhhhccccchhhhhhhhhccc

 SEQ TLGVETCGSKGNAEKSMSATEKQERTVMSHGCENFNTRDSVTGKESQGERPHLSLFIPRE
 SEG
35 PRD eeeeccccccccchhhhhhhhhhhhhhhccccccccccccccccccccccccceeecccc

 SEQ TTYQFHSVSQSSSQSLASLATTFLQEKKAQAQNHNCVPDVKALMESPEGQLSLEPKSDSQ
 SEG xxxxxxxxxxxx.....
40 PRD eeeeeeccccccccchhhhhhhhhhhhhhhhhhhhhccccccccchhhhhcccccccccccccccc

 SEQ FQASHTGCQSP LCSAQ RHTPQSPFTNHAAAAGRKTLRSCMGLEWFP ELYPGYLGLGVLP G
 SEG xxxxxxxxxxxxxxxxxxxx
 PRD ccccccccccccccccccccccccccccchhhhhcchhhhhccccccccccccccccceeeccc

45 SEQ KPQCWNAMTQKPQLISPQGERLSQGWIRCNTTIRKSGFGGITMLFTGYFVLCCSWSFRR L
 SEG xx.....
 PRD ccccccccccccccccccccccccchhhhhccccceeeccccccccceeeccccccccchhhhh

 SEQ KKLCRPLPWKSTVPPCIGVAKTTGD CRSKTCLD
50 SEG
 PRD hhccccccccccccccccceeecccccccccccccc

(No Prosite data available for DKFZphtes3_11e17.3)

(No Pfam data available for DKFZphtes3_11e17.3)

5 group: testis derived

DKFZphtes3_l2d18 encodes a novel 1170 amino acid protein without similarity to known proteins.

10 The EST-distribution signifies an ubiquitous expression pattern. No informative BLAST results; No predictive prosite, pfam or SCOP motife.

15 The new protein can find application in studying the expression profile of testis-specific genes.

unknown protein

20 perhaps complete cds.

Sequenced by Qiagen

Locus: /map="13b.9 cR from top of Chr13 linkage group"

25

Insert length: 5469 bp

Poly A stretch at pos. 5449, polyadenylation signal at pos. 5420

```

30      1 AAGGACAGAG GACGAGATTT TGAACGACAA AGAGAAAAGA GAGACAAGCC
      51 AAGGTCTACT TCCCCAGCAG GACAGCATCA TTCTCCTATA TCTTCTAGAC
     101 ATCACTCATC TTCCTCACAA TCAGGATCAT CTATTCAAAG ACATTCTCCT
     151 TCTCCTCGTC GAAAAAGAAC TCCTTCACCA TCTTATCAGC GGACACTAAC
     201 TCCACCTTTA CGACGCTCTG CCTCTCCTTA TCCTTCACAT TCTTTGTCGT
35     251 CTCCCCAGAG AAAGCAGAGT CCTCCAAGAC ATCGCTCTCC AATGCGAGAG
     301 AAAGGGAGAG ATGATCATGA ACGAACTTCA CAGTCTCATG ATCGACGCCA
     351 CGAAAGGAGG GAAGATACTA GGGGCAAACG AGACAGAGAA AAGGACTCAA
     401 GAGAAGAACG AGAATATGAA CAGGATCAGA GCTCTTCTAG AGACCACAGA
     451 GATGACAGAG AACCTCGAGA TGGTCGGGAT CGGAGAGATG CCAGAGATAC
40     501 TAGGGACCGA AGGGAACATA GAGACTCCAG AGACATGCGG GACTCAAGGG
     551 AGATGAGAGA TTATAGCAGA GATACCAAAG AGAGCCGTGA TCCCAGAGAT
     601 TCTCGGTCCA CTCGTGATGC CCATGACTAC AGGGACCGTG AAGGTCGAGA
     651 TACTCATCGA AAGGAGGATA CATATCCAGA AGAATCCCGG AGTTATGGCC
     701 GAAACCATTT GAGAGAAGAA AGTTCTCGTA CGGAAATAAG GAATGAGTCC
45     751 AGAAATGAGT CTCGAAGTGA AATTAGAAAT GACCGAATGG GCCGAAGTAG
     801 GGGGAGGGTT CCTGAGTTAC CTGAAAAGGG AAGTCGAGGC TCAAGAGGTT
     851 CTCAAATTGA TAGTCACAGT AGTAATAGCA ACTATCATGA CAGCTGGGAA
     901 ACTCGAAGTA GCTATCCTGA AAGAGATAGA TATCCTGAAA GAGACAACAG
     951 AGATCAAGCA AGGGATTCTT CCTTTGAGAG AAGACATGGA GAGCGAGACC
50    1001 GTCGTGACAA CAGAGAGAGA GATCAAAGAC CAAGCTCACC AATTTCGACAT
     1051 CAGGGAAGGA ATGACGAGCT TGAGCGTGAT GAAAGAAGAG AGGAACGAAG
     1101 AGTAGACAGA GTGGATGATA GGAGAGATGA AAGGGCTAGA GAGAGAGATC
     1151 GGGAACGAGA ACGAGACAGG GAGCGGGAGA GAGAGAGGGA ACGTGAACGG
     1201 GATCGGGAAA GAGAAAAAGA GAGAGAACTA GAAAGAGAGC GTGCTAGGGA
55    1251 ACGGGAGAGA GAAAGAGAAA AAGAGAGAGA TCGTGAAAGG GATAGAGACC
     1301 GAGACCACGA TCGAGAGCGG GAAAGAGAGA GGGAACGAGA CAGGGAAAAA
     1351 GAACGGGAAC GAGAAAGAGA AGAGAGAGAG AGGGAGAGAG AGCGAGAACG
     1401 GGAGAGAGAG CGAGAGCGAG AACGGGAACG AGAAAGAGCG AGAGAAAGGG

```

	1451	ATAAAGAACG	AGAACGCCAA	AGGGATTGGG	AAGACAAAGA	CAAAGGACGA
	1501	GATGACCGCA	GAGAAAAGCG	AGAAGAGATC	CGAGAAGATA	GGAAATCCAAG
	1551	AGATGGACAT	GATGAAAGAA	AATCAAAGAA	GCGCTATAGA	AATGAAGGGA
	1601	GTCCCAGCCC	TAGACAGTCC	CCGAAGCGCC	GGCGTGAACA	TTCTCCGGAC
5	1651	AGTGATGCCT	ACAACAGTGG	AGATGATAAA	AATGAAAAAC	ACAGACTCTT
	1701	GAGCCAAGTT	GTACGACCTC	AAGAATCTCG	TTCTCTTAGT	CCCTCGCACC
	1751	TCACAGAAGA	CAGACAGGGT	AGATGGAAAG	AGGAGGATCG	TAAACCAGAA
	1801	AGGAAAGAGA	GTTCAAGGCG	CTACGAAGAA	CAGGAACTCA	AGGAGAAAAGT
	1851	TTCTTCTGTA	GATAAACAGA	GAGAACAGAC	AGAAATCCTG	GAAAGCTCAA
10	1901	GAATGCGTGC	ACAGGACATT	ATAGGACACC	ACCAGTCTGA	AGATCGAGAG
	1951	ACATCTGATC	GAGCTCATGA	TGAAAAACAAG	AAGAAAGCAA	AAATTCAAAA
	2001	GAAACCAATT	AAGAAAAAGA	AAGAGGATGA	TGTTGGAATA	GAGAGGGGTA
	2051	ACATAGAGAC	AACATCTGAA	GATGGTCAAG	TATTTTCACC	AAAAAAAGGA
	2101	CAGAAAAAGA	AAAGCATTGA	AAAAAAACGT	AAAAAATCCA	AAGGTGATTG
15	2151	TGATATTTCT	GATGAAGAAG	CAGCCCAGCA	AAGTAAGAAG	AAAAGAGGCC
	2201	CACGGACTCC	CCCTATAACA	ACTAAAGAGG	AATTGGTTGA	AATGTGCAAT
	2251	GGTAAGAATG	GTATTCTAGA	GGACTCCAG	AAAAAAGAAG	ATACAGCATT
	2301	CAGTGACTGG	TCTGATGAGG	ATGTCCCTGA	CCGTACAGAG	GTGACAGAAG
	2351	CAGAGCATAAC	TGCCACCGCC	ACGACTCCTG	GTAGTACCCC	TTCTCCTCTA
20	2401	TCTTCTCTTC	TTCTCTCTCC	ACCGCTGTG	GCTACTGCCA	CTGCTACAAC
	2451	TGTGCTTGCA	ACTCTTGCTG	CCACTACTGC	TGCTGCCGCC	ACCTCTTTCA
	2501	GCACATCTGC	CATCACTATT	TCCACCTCTG	CCACCCCCAC	CAATACCACC
	2551	AATAATACTT	TTGCCAATGA	AGACTCACAC	AGAAAATGCC	ACAGAACACG
	2601	AGTAGAAAAA	GTAGAGACGC	CTCACGTGAC	TATAGAAGAT	GCACAGCATC
25	2651	GCAAGCCTAT	GGATCAAAAAG	AGGAGCAGCA	GCCTCGGGAG	CAATCGGAGT
	2701	AACCGTAGTC	ATACGTCTGG	TCGTCTTCGC	TCCCCATCCA	ATGATTGAGC
	2751	CCATCGAAGT	GGAGATGACC	AAAGTGGTCC	AAAGAGAGTA	CTGCACAGTG
	2801	GCTCAAGAGA	TAGAGAAAAA	ACAAAAAGCC	TGGAAATCAC	AGGAGAGAGA
	2851	AAATCTAGGA	TTGATCAGTT	AAAGCGTGGG	GAACCCAGTC	GAAGTACTTC
30	2901	TTCAGATCGC	CAGGATTCAA	GAAGCCATAG	TTCAAGAAGA	AGTTCTCCAG
	2951	AGTCAGATCG	ACAGGTCCAT	TCAAGATCTG	GGTCATTTGA	TAGCAGAGAC
	3001	AGGCTTCAAG	AACGAGATCG	ATATGAACAC	GACAGAGAGC	GCGAGAGAGA
	3051	GAGGAGAGAT	ACGAGGCAGA	GAGAATGGGA	CCGAGATGCT	GATAAAGATT
	3101	GGCCACGCAA	CAGGGATCGA	GATAGATTGC	GAGAACGAGA	ACGAGAGAGA
35	3151	GAACGAGACA	AAAGGAGAGA	CTTGGATAGG	GAAAGAGAGA	GACTAATTTT
	3201	TGATTCTGTT	GAAAGGGAGA	GGGACAGAGA	CAGAGACAGA	ACTTTTGAGA
	3251	TTTCTCAAAT	AGAGTCTGTG	AAACGCTGTG	AAGCAAAACT	GGAAGGTGAA
	3301	CATGAAAGGG	ATCTAGAAAG	CACCTCCCGA	GACTCTCTAG	CCTTGGATAA
	3351	AGAGAGAATG	GATAAAGATC	TGGGATCTGT	GCAGGGATTT	GAAGATACAA
40	3401	ATAAATCCGA	GAGAACTGAG	AGTCTGGAAG	CAGGAGATGA	CGAGTCCAAG
	3451	TTAGATGATG	CACATTTCATT	AGGCTCTGGT	GCTGGAGAAG	GATACGAGCC
	3501	AATCAGTGAT	GACGAACTAG	ATGAAATTCT	GGCAGGTGAT	GCAGAAAAGA
	3551	GGGAGGACCA	ACAGGATGAG	GAGAAGATGC	CAGATCCCTT	AGATGTGATA
	3601	GATGTGGATT	GGTCTGGTCT	TATGCCAAAG	CATCCAAAAG	AACCACGAGA
45	3651	GCCTGGGGCT	GCACTCTTAA	AATTCACACC	TGGAGCTGTT	ATGCTAAGAG
	3701	TTGGGATTTT	TAAAAAGTTG	GCAGGTTCTG	AACTCTTTGC	CAAAGTCAAA
	3751	GAAACATGTC	AGAGACTTTT	AGAAAAACCC	AAAGGTAGTT	TCATTTTACT
	3801	TTAACTATAT	AATGTCTGTT	AACCATTTAA	GATGCCATCT	GAAGGGGATT
	3851	CTGATCTGTT	CTTATGTAGC	ACTTAACACT	GTGTAGAAAC	TATTTTTTGA
50	3901	GAAATCATT	TATAATCATT	ATTTAACCT	CATGGTCAAA	GTTTCTCTTT
	3951	AAAATTTATT	TTGAGAAGAA	GAGTTATCCC	ACAGAAAAGT	TGGGAAAAGA
	4001	GTACAATGAC	CTTTTTGTAT	GAAAATTACT	TATTAACAGG	CCAGGCGTGG
	4051	TGTTGCATGT	CTGTAGTCAC	AGTACTCAG	GGAGGTTGAG	GCAGCAGGAT
	4101	TGCTGGAGCC	CAGGAAATTG	AGGCTGCAGT	GAGCCATGAT	TGAGCCACCA
55	4151	CACTCCAACC	TAGGTGACAG	AGCAAGACCC	TGTCTCAAAA	AAAAAAAAC
	4201	AAATTAACCA	ATAAGTTCTA	ATATCAAAGT	GCTCAGTGGT	TTGCCCTTGG
	4251	CTAAATGAAG	CAGAGCCAGG	AAAAACAGAC	TACATATTTT	TCATGTCTAA
	4301	AGAAATTGGG	TATTTTGGCA	GCCCTTTCCC	CTAGACATCT	ACCCAAATGC

```

4351 AGGTGTGTAG GTTGAGTCTT TAACAAAGTG ATTAAGAGCT TGGTCTGTAA
4401 GGCCGGATGA TCTGGATTTT AGTAGGCACA CCACTTACTG GCTATTACTT
4451 AATCTGTGTG TTAGTGTCAT CATCTGTAAG TCAGGAATAA TCATACCACC
4501 AACTTCCTAT GGTAATTAGG AGCAAATGAG TTATTACAGG CAAAACACTT
5 4551 AGAACAGTTC CTGGCATATA GTAATACCCA ATAAATATTA ACTGCTACTT
4601 TGAAAATATC CTATCACGCT GATTTTTTGAC CTCACTGCAG CAATTTTCAG
4651 TTATTCCAGA TTATCTAGCT TATGGATTCT GGTGGTAGGG GTTGTGTTGGT
4701 TTTGGTTTTT ACTGTCTCTG TCTCATCTAG TACCTACCTT AGTTTATTTT
4751 GCAACTTACT AATACTTTAT TAATGGGGAG GGACGAGTAG ATGGTAAAAA
10 4801 GAAGGAAAAG GAGGTAAAAG GTGAAAGGAA CAACATTAAT TAACAATTTT
4851 ACGTCATGTC CCTGGACATA AAAGTTTAGT TAGTATTAAT TTTTTCATA
4901 ATACAAAATA AAAAAATATT GTTTTATGAG TTTTATGAAT TCATGCCCTT
4951 CCTTACTCT ATTAGCATAA GCAGTAAATT TTTTATTTT AATATAGCCC
5001 AATAAACCTA GAGTATACAT GTACAAAATA CATATAATTG TTAACGTGTA
15 5051 TTAACCGAAA AATGACCCAA GACTTAGTTC TTGCCCTACT GTATCTGCCT
5101 TGTTTGTTG GTTCTGTGAC CTTAAGCAAA TAACTCCTGT GAGCCTCAAT
5151 TTTATTTGTA AAGTGATGGA ATAAAACCCC TAAAATCTTA CCCACCTCTA
5201 AAGATATTTG TTTCTGTGAC CTTTTGCTAG TAGCATTTCA AGTTAAAATC
5251 TGGTTTGATT TTGCTACCCA TGAAATACAG TTCGGCCCTT ACTTATTGAT
20 5301 GACTTAACCT AAACAGTGAA AATATGCACT GTAAAGGGTG GGGTGATGTG
5351 GCTTAACAAT CAGACTTCTT CTATTTTGC TGCTATGGTG GTTGTATTAG
5401 AGAACTGATG TATTATCTTG AATAAAGACT TTGTCTTGTT TACTGCCCTA
5451 AAAAAAAAAA AAAAAAAAAA

```

BLAST Results

No BLAST result

Medline entries

No Medline entry

Peptide information for frame 1

ORF from 292 bp to 3801 bp; peptide length: 1170
Category: similarity to unknown protein
Classification: no clue

```

1 MREKGRHDHE RTSQSHDRRH ERREDTRGKR DREKDSREER EYEQDQSSSR
51 DHRDDREPRD GRDRRDARDT RDRRELDSR DMRDSREMRD YSRDTKESRD
101 PRDSRSTRDA HDYRDREGRD THRKEDTYPE ESRSYGRNHL REESSRTEIR
151 NESRNESRSE IRNDRMGRSR GRVPELPEKG SRGSRGSQID SHSSNSNYHD
50 201 SWETRSSYPE RDRYPERDNR DQARDSSFER RHGERDRRDN RERDQRPSSP
251 IRHQGRNDEL ERDERREERR VDRVDDRRDE RARERDRERE RDRERERERE
301 RERDREREKE RELERERARE REREREKERD RERDRDRDHD RERERERERD
351 REKERERERE ERERERERER ERERERERER ERARERDKER ERQRDWEDKD
401 KGRDDRREKR EEIREDRNPR DGHDERKSKK RYRNEGSPSP RQSPKRRREH
55 451 SPDSDAYNSG DDKNEKHRL L SQVVRPQESR SLSPSHLTED RQGRWKEEDR
501 KPERKESSRR YEEQELKEKV SSVDKQREQT EILESSRMRA QDIIGHHQSE
551 DRETSDRAHD ENKKKAKIQK KPIKKKKEDD VGIERGNIET TSEDGQVFSP
601 KKGQKKKSIE KKRKKS KGDS DISDEEAAQ SKKKRGPRTP PITTKELVE

```

```

5  651 MCNGKNGILE DSQKKEDTAF SDWSDDEVDP RTEVTEAEHT ATATTPGSTP
    701 SPLSSLLPPP PPVATATATT VPATLAATTA AAATSFSTSA ITISTSATPT
    751 NTTNNTFANE DSHRKCHRTR VEKVETPHVT IEDAQHRKPM DQKRSSSLGS
    801 NRSNRSHTSG RLRSPSND SA HRSRDDQSGR KRVLHSGSRD REKTKSLEIT
10  851 GERKSRIQQL KRGEPSRSTS SDRQDSRSHS SRRSSPESDR QVHSRSGSFD
    901 SRDRLQERDR YEHDREERERE RRDTRQREW DADADKDWPRN RDRDRLRERE
    951 RERERDKRRD LDRERERLIS DSVERDRDRD RDRTFESSQI ESVKRCEAKL
   1001 EGEHERDLES TSRDSLALDK ERMDKDLGSV QGFEDTNKSE RTESEAGDD
   1051 ESKLDDAHS L GSGAGEGYEP ISDDELDEIL AGDAEKREDQ QDEEKMPDPL
10  1101 DVIDVDW SGL MPKHPKEPRE PGAALLKFTP GAVMLRVGIS KKLAGESELF A
   1151 KVKETCQRL L EKPKGSFILL

```

15 BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_12d18, frame 1

20 No Alert BLASTP hits found

Pedant information for DKFZphtes3_12d18, frame 1

25 Report for DKFZphtes3_12d18.1

```

30  [LENGTH] 1267
    [MW] 150593.45
    [pI] 9.22
    [HOMOL] TREMBL:AB020660_1 gene: "KIAA0853"; product:
    "KIAA0853 protein"; Homo sapiens mRNA for KIAA0853 protein,
    partial cds. 0.0
35  [BLOCKS] BL00422C Granins proteins
    [BLOCKS] BL00803F
    [BLOCKS] PRO0308C
    [BLOCKS] PRO1089B
    [BLOCKS] PRO0049D
40  [BLOCKS] PRO1083A
    [BLOCKS] PRO0545A
    [BLOCKS] BL00048 Protamine P1 proteins
    [BLOCKS] PF01140D
    [BLOCKS] PRO0833H
45  [KW] All_Alpha
    [KW] LOW_COMPLEXITY 44.12 %

```

```

50  SEQ KDRGRD FERQREKRDKPRSTSPAGQHHSPISSRHSSSSQSGSSIQRHSPSPRRKRTPSP
    SEG .....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
    PRD cccccchhhhhhhcccccccccccccccccccccccccccccccccccccccccc

    SEQ SYQRTLTPLRRSASPYP SHSLSSPQRKQSPPRHRSPMREKGRHDHERTSQSHDRRHERR
    SEG x.....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
55  PRD cccccccccccccccccccccccccccccccccccccccccccccccccccccchhhhhhc

    SEQ EDTRGKRDREKDSREEREYEQDQSSSRDHRDDREPRDGRDRRDARDTRDRREL RDSRDMR
    SEG xx.xxxxxxxxxxxxxxxxxxxxxx.xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

```

PRD cccccccccccchhhhhhhhhccccccccccccccccccccchhhhhhhhhhhhhhhhhcccc

5 SEQ DSREMRDYSRDTKESRDPRDSRSTRDAHDYRDREGRDTHRKEDTYPEESRSYGRNHLREE
SEG xxxxxxxxxxxx...xxxxxxxxxxxxx.....
PRD hhhhhhhcc

10 SEQ SSRTEIRNESRNESRSEIRNDMGRSRGRVPELPEKGSRGSRGSQIDSHSSNSNYHDSWE
SEGxxxxxxxxxxxxx.:
PRD hhhhhhhcc

15 SEQ TRSSYPERDRYPERDNRDQARDSSFERRHGERDRRDNRERDQRPSSPIRHQGRNDELERD
SEGxxxxxxxxxxxxxxxxxxxxx.....xxxxxx
PRD cchhhhhh

20 SEQ ERREERRVDRVDDRRDERARERDRERERDRERERERERERDREREKERELERERARERER
SEG xxx
PRD hhhhhhhhhccccccccchhh

25 SEQ EREKERDRERDRDRDHDRERERERERDREREKEREREREERERERERERERERERERERERA
SEG xxx
PRD hhhhhhhhhccccccccchhh

30 SEQ RERDKERERQRDWEDKDKGRDDRREKREEIREDNRPRDGHDERKSKKRYRNEGSPSPRQS
SEG xxxxxxxx..xxxxxxxxxxxxxxxxxxxxx.....xxxxxx
PRD hhhhhhhhhhhhhhhhhccccccccchhhhhhhhhhhccccccccccccchhhhhcccccccccc

35 SEQ PKRRREHSPDSDAYNSGDDKNEKHRLLSQVVRPQESRSLSPSHLTEDRQGRWKEEDRKPE
SEG xxxxx.....
PRD cccccccccccccccccccccchhhhhhhhhccccccccccccccccccccchhhhhhhhhhhccch

40 SEQ RKESSRRYEEQELKEKVSSVDKQREQTEILESSRMRAQDIIGHHQSEDRETSdrahdenk
SEGx
PRD hhhhhhhhhhhhhhhhhhhccchhhhhhhhhhhhhhhhhheeeccccccccccccccccccch

45 SEQ KKAKIQKKPIKKKKEDDVGIERGNIETTSEDGQVFSPPKKGQKKKSIEKKRKSKGDSDIS
SEG xxxxxxxxxxxxxxxx.....xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
PRD hhhhhhhhhccccccccccccccccccccccccccccccccccccchhhhhhhhhhhcccccccc

50 SEQ DEEAAQQSKKKRGPRTPPITTKELVEMCNGKNGILEDQKKEDTAFSDWSDVDVPDRTE
SEG xxx.....xx
PRD hhhhhhhhhhhccccccccccccchhhhhhhcccccccccccccccccccccccccccccccc

55 SEQ VTEAEHTATATTPGSTPSPLSSLLPPPPPVATATATTVPATLAATTAATAATSFSTSAITI
SEG xxx
PRD hhhhhhhhhccccccccccccccccccccccccccccccccccccchhhhhhhhhhhhhhhcccccccc

60 SEQ STSATPTNTTNTFANEDSHRKCHRTVEKVPHTVIEDAQHRKPMQKRSSSLGNSRS
SEG xxxxxxxxxxxxxxxx.....xxxxxxxxxxxx
PRD eccccccccccccccccccccccccchhhhhheeecccccccccccccccccccccccccccccccc

65 SEQ NRSHTSGRLRSPSND SAHRSQDDQSGRKRVLHSGSRDREKTKSLEITGERKSRIDQLKRG
SEG xxx.....
PRD cchhhhhhhhhhhhhcc

70 SEQ EPSRSTSSDRQDSRSHSSRRSSPESDRQVHSRSGSFDSRDRLQERDRYEHDRERERERRD
SEG ..xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.....xxxxxxxxxxxxxxxxxxxxxx
PRD cchhhhhhhhhhhchhhhhhhhhhh

20

(No Pfam data available for DKFZphtes3_l2d18.1)

• 25 •

5. group: testis derived
- DKFZphtes3_1417 encodes a novel 815 amino acid protein without similarity to known proteins.
- 10 The mRNA is transcribed ubiquitously.
No informative BLAST results; No predictive prosite, pfam or SCOP motif.
- 15 The new protein can find application in studying the expression profile of testis-specific genes.

similarity to C.elegans B0412.3

- 20 see also DKFZphtes3_17n3
perhaps complete cds.

Sequenced by BMFZ

- 25 Locus: unknown

Insert length: 3522 bp
Poly A stretch at pos. 3456, polyadenylation signal at pos. 3437

```

30      1 AACACATCGA CTTGTGTAAG AAAAAGATTG GAAGTGCGGA GCTGTCTTTT
      51 GAGCATGATG CATGGATGTC TAAACAATTC CAGGCCTTTG GAGATTTATT
     101 TGATGAAGCT ATTAAGTTAG GGTTAACAGC TATTCAAACCT CAGAATCCTG
     151 GTTTCTATTA CCAGCAGGCA GCATACTATG CCCAGGAGCG GAAACAGCTT
    201 GCAAAAACCC TCTGTAACCA CGAAGCTTCT GTAATGTATC CCAATCCTGA
    251 TCCCTTAGAA ACACAAACAG GCGTTCTTGA CTTTTATGGA CAAAGATCAT
    301 GCGGACAAGG AATACTAAGT TTTGATCTTT CTGATCCTGA AAAAGAAAAG
    351 GTGGGAATTG TTGCCATTCA GCTGAAGGAG AGAAATGTTG TTCACTCTGA
    401 GATAATCATA ACTCTTCTGA GCAATGCTGT TGCACAGTTC AAGAAGTATA
    451 AGTGCCCGCG AATGAAAAGT CACCTAATGG TTCAGATGGG AGAGGAATAT
    501 TATTACGCAA AGGATTATAC CAAAGCTTTG AAGTTGCTGG ATTATGTGAT
    551 GTGTGATTAT CGGAGTGAAG GATGGTGGAC TCTGCTCACT TCTGTATTAA
    601 CTACAGCTCT GAAGTGCTCC TACCTCATGG CCCAATTAAA GGATTACATT
    651 ACTTACTCCC TAGAACTCCT TGGTAGAGCT TCAACTCTGA AAGATGACCA
    701 GAAGTCTCGG ATAGAAAAGA ACCTCATAAA TGTTTTAATG AATGAAAGTC
    751 CTGATCCAGA ACCCGACTGT GATATCTTAG CTGTGAAAAC TGCTCAGAAG
    801 CTGTGGGCAG ACCGAATTTT TCTGGCTGGC AGCAATATTT TCACAATAGG
    851 AGTACAGGAC TTTGTGCCAT TTGTGCAGTG CAAAGCCAAG TTTCATGCCC
    901 CAAGTTTTCA TGTTGATGTT CCTGTTCACT TTGATATTTA TCTGAAGGCT
    951 GATTGTCCAC ATCCCATTAG GTTTTCCAAG CTCTGTGTCA GCTTTAATAA
   1001 TCAGGAATAC AACCAGTTCT GTGTAATAGA AGAAGCATCC AAAGCAAATG
   1051 AAGTTTTAGA AAATCTGACT CAAGGAAAAG TGTGCCTAGT TCCTGGCAAA
   1101 ACAAGAAAAC TGTTATTTAA GTTTGTTGCA AAAACTGAAG ATGTGGGAAA
   1151 GAAAATTGAG ATTACTTCAG TGGATCTTGC TCTGGGCAAT GAGACGGGAA
   1201 GATGTGTGGT TTTAAATTGG CAGGGAGGAG GAGGAGATGC TGCTTCCTCC
   1251 CAAGAAGCCT TACAGGCAGC TCGGTCTTTC AAAAGGCGAC CTAAGCTACC
   1301 TGACAATGAA GTTCACTGGG ACAGCATTAT AATTCAGGCA AGCACAATGA
   1351 TCATATCCAG AGTCCCAAAC ATTTCTGTAC ATCTGCTACA TGAACCCCT

```

1401 GCACTGACTA ATGAAATGTA TTGTTTGGTT GTGACTGTTC AGTCCCATGA
1451 AAAGACCCAA ATCAGAGATG TGAAGCTCAC TGCTGGCTTA AAACCAGGAC
1501 AGGATGCCAA TTAACTCAG AAGACTCACG TGAATCTTCA TGGACCAGAA
1551 CTGTGTGATG AATCCTACCC GGCTTTACTC ACTGACATTC CTGTTGGAGA
5 1601 CTTACATCCA GGGGAACAGC TGGAAAAAAT GTTGTATGTT CGCTGTGGAA
1651 CAGTGGGTTC CAGAATGTTT CTTGTATATG TTTCTTACCT GATAAATACA
1701 ACCGTTGAAG AAAAAGAAAT TGTTTGCAAG TGTACAAGG ATGAAACTGT
1751 AACAAATTGAA ACAGTCTTTC CATTTGATGT TGCAGTTAAA TTTGTTTCTA
1801 CCAAGTTTGA GCACCTGGAA AGGGTTTATG CTGACATCCC CTTTCTGTTG
10 1851 ATGACGGACC TCTTAAGTGC CTCACCCTGG GCCCTCACTA TTGTTTCCAG
1901 TGAGCTCCAG CTTGCTCCAT CCATGACCAC AGTGGACCAG CTCGAGTCTC
1951 AAGTGGACAA TGTTATCTTA CAGACTGGAG AGAGTGCTAG TGAATGCTTT
2001 TGTCTTCAAT GCCCATCTCT TGGAAATATT GAAGGTGGAG TAGCAACCGG
2051 GCATTATATT ATCTCTTGGA AAAGGACCTC AGCAATGGAG AATATCCCCA
15 2101 TCATCACAAC TGTCATCACT CTGCCGCACG TGATTGTGGA GAATATCCCT
2151 CTCCATGTGA ATGCAGATCT GCCGTCATTT GGGCGTGTCA GAGAGTCGTT
2201 ACCTGTCAAG TATCACCTAC AGAATAAGAC CGACTTAGTT CAAGATGTAG
2251 AAATTTCTGT GGAGCCCAAG GATGCCTTCA TGTCTCAGG TCTCAAACAG
2301 ATTCGATTAC GTATCCTCCC TGGCACGGAG CAGGAAATGC TATATAATTT
20 2351 CTATCCTCTG ATGGCTGGAT ACCAGCAGCT GCCATCTCTC AACATCAACT
2401 TGCTTAGATT TCCTAACTTC ACAAATCAGC TGCTCAGGCG TTTTATACCT
2451 ACCAGTATTT TTGTCAAGCC ACAGGGTCGA CTCATGGATG ATACCTGAT
2501 TGCTGCTGCA TGATGTTCAA GACCGGCCCT TGGCTGTTGT TACAGAGATG
2551 TTGGGCAGAG CTATGCAGGT GTTTCATTGT GAACTCTAGC TTTGATCATG
25 2601 GTAAAAAGTT AACCTTTTCT ATTTTTTAAT GGATGTTATA CCAACTATTC
2651 AGAGGAACTC AACTTTCAAA AATATTAGGA AAATCTGTCT TATAGTTTCT
2701 CTAATAAATA TCTGAAATCT CAGTACGACA TGAAAGAATG TCAGACCATT
2751 GTTATTGTTG AAAGTCATTT GATGAATGGT AAATCTATG AAAAGTAAAGT
2801 GATTTGCATG TATAATATCA GGGAAATTAAG GCATCCCAAG TGTGACTGGA
30 2851 CAAAGAGAGC AGATGCACCA GTGCCTGTGC CATAAAGTTC CGAATCCCCC
2901 ATGTGTCTCT TTCAGAGCTG GCCAGACCGG AAATAAATCA TTCTCATAAA
2951 TTCAGTGTGT ACTCAGAACA CATAACAAC AACATAGGGA GTTGTATGAC
3001 TGATACGGAA AACTTCCAGA AAGTTTTAAT CAAAGCAGTT TAATTAAGGT
3051 ATCAAAAATA TCTTTGCTTA CTATCAAGAA GTGTCAAATA GGTTCAGCTT
35 3101 GCTGCCAAAA TATGGATCAT TTATGAAGCA GGTTCATATT TTAGAGGTGT
3151 TAATAAAATC CTCATGGGAA AAGATCCAAA GTGCAAGGAT TTGATTATAA
3201 ACATAATTTT CTAGACTGAA AGTTTTTGGG AAAGATGCAG GGTCTGAGTC
3251 AGGCCTTCTG GTTATATTGT GCAGTTTCAA AAGAACTATT TAAAACCTTT
3301 GAAAACATCAT GTAAATAAAA ATCATAGGGT GAAAATTGTA TTTGTTAAAA
40 3351 TACCTTAATA ATTTAAATG ACCTGATTTC CTGGAAAATT TTATTATTCA
3401 AAAGGTGGAG GCATTGTAAA AAGGAAATAG TGATGTAAAT AAACATGTTT
3451 TCTTTCAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
3501 AAAAAAAAAA AAAAAAAAAA AA

45

BLAST Results

No BLAST result

50

Medline entries

55 No Medline entry

5 Category: similarity to unknown protein
Classification: no clue

BLASTP hits

Alert BLASTP hits for DKFZphtes3_1417, frame 3

35 Pedant information for DKFZphtes3_1417, frame 3

40

```

45  [HOMOL]      TREMBL:CEUB0412_2 gene: "B0412.3"; Caenorhabditis
    elegans cosmid B0412. 6e-30
    [KW]        Alpha_Beta
    [KW]        LOW_COMPLEXITY      1-20 %

```

-308-

(No Pfam data available for DKFZphtes3_1417.3)

DKFZphtes3_15n14

5 group: testis derived

DKFZphtes3_15n14 encodes a novel 713 amino acid protein with weak similarity to the neurofilament triplet M protein of the rat.

10 Neurofilaments are the intermediate filaments specific to nervous tissue. They are probably essential to the tensile strength of the neuron, as well as to transport of molecules and organelles within the axon. Until now, ESTs of the novel mRNA could only be isolated from testes, germ cells and uterus.

15 No informative BLAST results; No predictive prosite, pfam or SCOP motife.

The new protein can find application in studying the expression profile of testis-specific genes.

20

similarity to neurofilament triplet M protein - rat

few EST hits (6 of 9 hits from testis)

25 perhaps complete cds.

Sequenced by GBF

Locus: unknown

30

Insert length: 2389 bp

Poly A stretch at pos. 2328, polyadenylation signal at pos. 2306

```

35      1 TGGGCCCCAC CTCCTCAGCA CAACTTTCTG AAAAAGTGGC AGCGTAACAC
      51 AGCCCTGCGG AAGAAGCAGC AGGAAGCCCT CAGCGAACAC CTAAAGAAGC
     101 CAGTGAGTGA GCTGCTCATG CACACCGGGG AGACCTACAG ACGGATCCAG
     151 GAGGAGCGGG AGCTCATTGA CTGCACACTT CCAACCCGGC GTGATAGGAA
     201 AAGCTGGGAG AACAGTGGGT TCTGGAGTCG ACTGGAATAC TTGGGAGATG
     40 251 AGATGACAGG TCTGGTCATG ACCAAGACAA AAAGCTCAGCG TGGCCTCATG
     301 GAGCCCATCA CTCATATCAG GAAGCCCCAC TCCATCCGGG TGGAGACAGG
     351 ATTACCAGCC CAGAGGGACG CTTCATACCG CTACACCTGG GATCGGAGTC
     401 TGTTTCTGAT CTACCGACGC AAGGAGCTGC AGAGAATCAT GGAAGAGCTG
     451 GATTTTCAGCG AGCAGGATAT TGATGGCCTG GAGGTGGTGG GCAAAGGGTG
     45 501 GCCCTTCTCG GCTGTTACTG TGGGAAGACTA CACAGTGTTT GAAAGAAGTC
     551 AGGGAAGCTC CTCTGAAGAC ACAACATACT TAGGCACATT GGCCAGTTCC
     601 TCTGATGTCT CCATGCCTAT TCTCGGCCCT TCTCTGCTGT TCTGTGGGAA
     651 GCCAGCTTGC TGGATCAGAG GCAGTAATCC ACAGGACAAG AGGCAGGTTG
     701 GGATTGCTGC TCACTTGACC TTTGAAACCC TAGAAGGCGA GAAAACCTCC
     50 751 TCAGAACTGA CTGTGGTCAA TAATGGCACC GTGGCCATTT GGTATGACTG
     801 GCGACGGCAG CACCAGCCGG ACACTTTCCA AGACCTTAAG AAAAACAGGA
     851 TGCAGCGATT TTAATTTGAC AACCAGGGAAG GTGTGATTCT GCCTGGAGAA
     901 ATTAATAACAT TTACCTTCTT CTTCAAGTCT TTGACTGCTG GGGTCTTCAG
     951 GGAATTTTGG GAGTTTCGAA CCCATCCTAC TCTATTAGGA GGTGCTATAC
     55 1001 TGCAGGTCAA TCTCCACGCG GTCTCCCTGA CCCAGGACGT TTTTGAGGAT
    1051 GAGAGGAAAG TACTGGAGAG CAAGCTGACT GCCCATGAGG CAGTCACCGT
    1101 CGTTCGCGAA GTGCTGCAGG AGCTGCTGAT GGGGGTCTTG ACCCCGGAGC
    1151 GCACACCATC ACCTGTGGAT GCCTATCTCA CCGAGGAAGA CTTGTTCCGG

```

1201 CACAGAAATC CTCCGCTGCA TTATGAGCAC CAAGTGGTGC AAAGCCTGCA
 1251 CCAACTGTGG CGCCAGTACA TGACCCTGCC CGCCAAGGCT GAGGAGGCCA
 1301 GGCCAGGGGA CAAGGAGCAC GTCAGCCCCA TAGCCACAGA GAAGGCTCT
 1351 GTGAATGCTG AGCTGTTACC ACGCTTTAGG AGCCCCATCT CCGAAACTCA
 5 1401 AGTCCCCGGG CCTGAGAACG AGGCCCTCAG GGAATCCGGG TCCAGAAAGG
 1451 CCAGAGTGGG GACCAAGAGT CCTCAGCGGA AGAGCATCAT GGAGGAGATC
 1501 CTGGTGGAGG AAAGCCAGAG TGTGGACAGC ACCAAGAGCC CCTGGGAGCC
 1551 GGATGGCCTT CCCCTGCTGG AGTGGAACTT CTGCTTGGAG GACTTCAGAA
 1601 AGGCACTGAT GGTGCTCCCT GATGAGAACC ACAGAGAGGA TGCCTTGATG
 10 1651 AGGCTCAACA AAGCAGCCCT GGAGCTGTGC CAGAAGCCAA GGCCATTGCA
 1701 GTCCAACCTC CTGCACCAGA TGTGTTTGCA GCTGTGGCGA GATGTGATTG
 1751 ACAGCCTGGT GGGCCATTCC ATGTGGCTGA GGTCTGTGCT GGGCCTGCCT
 1801 GAGAAGGAGA CCATCTATTT GAATGTGCCT GAAGAGCAAG ATCAAAAATC
 1851 ACCTCCTATC ATGGAAGTGA AGGTACCTGT GGGGAAAGCT GGGAAAGGAGG
 15 1901 AGCGGAAAGG AGCAGCCCAG GAAAAGAAGC AACTGGGGAT CAAAGACAAA
 1951 GAAGACAAGA AAGGAGCCAA GCTGCTCGGG AAAGAGGACC GTCCCAACAG
 2001 CAAGAAGCAC AAGGCAAGG ATGACAAGAA AGTCATAAAA TCTGCAAGTC
 2051 AGGACAGGTT TTCTTTGGAA GACCCTACCC CTGACATCAT CCTCTCTTCT
 2101 CAAGAACCCA TAGACCCCTT GGTTCATGGGG AAATACACCC AGAGGCTGCA
 20 2151 CAGTGAGGTC CGTGGGCTGC TGGACACCCT GGTGACCGAC CTGATGGTCC
 2201 TGGCTGATGA GCTCAGCCCC ATAAAGAATG TCGAGGAGGC TTTGCGCCTC
 2251 TGCAGGTGAC TCTCGGGCCC AAGCAACCTT CTGGAAAACG GGTTAATAAA
 2301 TAAATCAATA AAGAACCTTC AAGTTTCTAC TAAAAAATAA AAAAAAATAA
 2351 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA GGGCGGCCG

BLAST Results

30 No BLAST result

Medline entries

35 No Medline entry

40 Peptide information for frame 1

ORF from 118 bp to 2256 bp; peptide length: 713

Category: putative protein

45 Classification: Cell structure/motility

1 MHTGETYRRI QEERELIDCT LPTRRDKRSW ENSGFWSRLE YLGDEMTGLV
 51 MTKTKTQRGL MEPITHIRKP HSIRVETGLP AQRDASYRYT WDRSLFLIYR
 101 RKELQRIMEE LDFSQQDIDG LEVVGKGWPF SAVTVEDYTV FERSQGSSE
 50 151 DTTYLGTLAS SSDVSMPILG PSLLFCGKPA CWIRGSNPQD KRQVGIAAHL
 201 TFETLEGEKT SSELTVVNNG TVAIWYDWRR QHQPDTFQDL KKNRMQRIFY
 251 DNREGVILPG EIKTFTFFFK SLTAGVFREF WEFRTHTPLL GGAILQVNLH
 301 AVSLTQDVFE DERKVLESKL TAHEAVTVVR EVLQELLMGV LTPERTPSPV
 351 DAYLTEEDLF RHRNPPLHYE HQVVQSLHQL WRQYMTLPK AEEARPGDK
 55 401 HVSPIATEKA SVNAELLPRF RSPISETQVP RPENEALRES GSQKARVGTK
 451 SPQRKSMEE ILVEESPDVD STKSPWEPDG LPLLEWNLCL EDFRKAVMVL
 501 PDENHREDAL MRLNKALEL CQKPRPLQSN LLHQMCLQLW RDVIDSLVGH
 551 SMWLRSVLGL PEKETIYLVN PEEQDQKSP IMEVKVPVGK AGKEERKGAA

5

No BLASTP hits available

10

15

20

25

35

40

45

50

55

```
SEQ RHRNPPLHYEHQVVQSLHQLWRQYMTLPAKAEEARPGDKEHVSPiatekASVNAELLPRF
SEG .....
PRD cccccccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
```

SEQ RSPISSETQVPRPENEALRESGSQKARVGTKSPQRKSIMEEILVEESPDVDSTKSPWEPDG
SEG
PRD cccccccccccccchhhhhccccccccccccccccchhhhhhhhhhhcccccccccccccccc

5 SEQ LPLLEWNLCLEDFRKAVMVLPDENHREDALMRLNKALELCQKPRPLQSNLLHQMLQLW
SEG
PRD cccccchhhhhhhhhhhhhccccccccchhhhhhhhhhhhhhhhhhhccccccccchhhhhhhhhhhhh

10 SEQ RDVIDSLVGHSMWLRSVLGLPEKETIYLVNPEEQDQKSPPIMEVKVPVGKAGKEERKGAA
SEG
PRD hhhhhhhhccchhhhhccccccccceeeeeccccccccccccccccceeeeeccccchhhhhhhhh

15 SEQ QEKKQLGIKDKEDKKGAKLLGKEDRPNSKKHKAKDDKKVIKSASQDRFSLEDPTPDIIILS
SEGxxxxxxxxxxxxxxxxxxxxx.....
PRD hhhhhhccccccccchhhhhccccccccccccccccccccceeeeeccccccccccccccccceeee

20 SEQ SQEPIDPLVMGKYTQRLHSEVRGLLDTLVTDLMVLADELSPIKNVEEALRLCR
SEGxxxxxxxxxxxxx.....
PRD cccccccceechhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccccchhhhhhhhhccc

(No Prosite data available for DKFZphtes3_15n14.1)

(No Pfam data available for DKFZphtes3_15n14.1)

DKFZphtes3_1bb5

5 group: cell structure and motility

DKFZphtes3_1bb5 encodes a novel 268 amino acid protein with similarity to various tropomyosins.

10 Tropomyosins play regulatory roles in cellular structure and transport.

The new protein can find application in modulating cell structure and motility as well as modulationg cellular transport.

15

weak similarity to KIAA0774

perhaps complete cds.

20

Sequenced by BMFZ

Locus: unknown

25 Insert length: 1316 bp

Poly A stretch at pos. 1247, polyadenylation signal at pos. 1232

```

      1 TGCTAAAATG GAATTAGAGA GAAGCATAGA CATCAGCAGA AGACAGAGTA
30    51 AGGAGCACAT ATGTAGAATT ACAGATCTAC AAGAGGAATT AAGACACAGA
      101 GAGCATCACA TCTCTGAATT GGATAAGGAG GTTCAGCACC TTCATGAGAA
      151 TATAAGTGCC CTAACCAAAG AACTGGAATT TAAGGGGAAA GAAATTCTCA
      201 GAATACGAAG TGAATCTAAC CAACAGATAA GGTTCATGA ACAAGATTTA
      251 AACAAGAGAC TTGAAAAAGA GTTGGATGTC ATGACAGCAG ACCACCTCAG
35    301 AGAGAAAAAT ATCATGCGGG CAGATTTTAA TAAGACTAAC GAGCTACTCA
      351 AGGAAATAAA TGCCGCTTTA CAAGTGTCAT TAGAAGAAAT GGAAGAAAAA
      401 TATCTAATGA GAGAATCAAA ACCAGAAGAT ATACAGATGA TTACAGAATT
      451 AAAAGCCATG CTTACAGAAA GAGACCAGAT CATAAAGAAA CTAATTGAGG
      501 ATAATAAGTT TTATCAGCTG GAATTAGTCA ATCGAGAAAC TAACCTCAAC
40    551 AAAGTGTTTA ACTCAAGTCC TACTGTTGGT GTTATTAATC CATTGGCTAA
      601 GCAAAAGAAG AAGAATGATA AATCACCAAC AAACAGGTTT GTGAGTGTTT
      651 CCAATCTAAG TGCTCTGGAA TCTGGTGGAG TGGGCAATGG ACATCCTAAC
      701 CGCCTGGATC CCATTCTTAA TTCTCCAGTC CACGATATTG AGTTCAACAG
      751 CAGCAAACCA CTTCCACAGC CAGTGCCACC TAAAGGGCCC AAGACATTTT
45    801 TGAGGTATCA GTAAGATGCA TGTGCATGAG CTCAAGGAAC ATGACTACTG
      851 GAGTTTCCAT TACACATTGT TGCCTGCCTT GTAATTTTCC CCAAAGACGT
      901 CCTGCTCAGA GTGAAGCTTC TCCAGTGGCT TCTCCAGATC CCCAGCGCCA
      951 GGAGTGGTTT GCCCAGTACT TCACATTCTG AAAGAATTGT GTTGGCACAG
50   1001 CTCTGTATAG ACTGTTACTA AGAGCATGAC TTTATACAGA TTGTTATGTA
      1051 AATAGGCTTT CCTATGTCAA ACACTGTGAA TGAGAAAGTA TTTGTCTCTC
      1101 CAACTTGAAA ATGCACTGTA TTTCTGTGA TATTTATTGG AATCATTCTA
      1151 TAAGGTACTA TATTATGTGT GTAATTATAA CTGTTATTTT TATTTGAGAT
      1201 GGAAGAGTCT TTAACCTTTG TAATTACTGC ATAATAAATT TTGTTAGAAT
      1251 CAAAAAATAA AAAAAAATAA AAAAAAATAA AAAAAAATAA AAAAAAATAA
55   1301 AAAAAAATAA AAAAAA

```

BLAST Results

No BLAST result

5

Medline entries

No Medline entry

10

Peptide information for frame 2

15

ORF from 8 bp to 811 bp; peptide length: 268
 Category: similarity to known protein
 Classification: Cellular transport and traffic

20

1 MELERSIDIS RRQSKHEICR ITDLQEELRH REHHISELDK EVQHLHENIS
 51 ALTKELEFKG KEILRIRSES NQQIRLHEQD LNKRLKELD VMTADHLREK
 101 NIMRADFNKT NELLKEINAA LQVSLEEMEE KYLMRESKPE DIQMITELKA
 151 MLTERDQIIK KLIEDNKFYQ LELVNRETNF NKVFNSSPTV GVINPLAKQK
 201 KKNDKSPTNR FVSVPNLSAL ESGGVGNHGP NRLDPIPNP VHDIEFNSSK
 251 PLPQPVPPKG PKTFLRYQ

25

BLASTP hits

30

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_16b5, frame 2

35

No Alert BLASTP hits found

Pedant information for DKFZphtes3_16b5, frame 2

40

Report for DKFZphtes3_16b5.2

45

[[LENGTH]] 270
 [[MW]] 31493.09
 [[pI]] 6.90
 [[HOMOL]] PIR:A57013 early endosome antigen 1 - human 1e-05
 [[FUNCAT]] 03.19 recombination and dna repair [[S. cerevisiae,
 Y0L034w]] 1e-05
 [[FUNCAT]] 03.22 cell cycle control and mitosis [[S. cerevisiae,
 YFR031c]] 2e-05
 [[FUNCAT]] 30.10 nuclear organization [[S. cerevisiae, YFR031c]]
 2e-05
 [[FUNCAT]] 11.04 dna repair (direct repair, base excision repair
 and nucleotide excision repair) [[S. cerevisiae, YKR095w]] 5e-05
 [[FUNCAT]] 30.04 organization of cytoskeleton [[S. cerevisiae,
 YDR356w]] 7e-05
 [[FUNCAT]] 09.10 nuclear biogenesis [[S. cerevisiae, YDR356w]]
 7e-05

55

```

    SEQ AKMELERSIDISRRQ$KEHICRITDLQEELRHREHHISELDKEVQHHLHENISALTKELEF
    SEG .....
20 PRD ccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
COILS .....CCCCCCCCCCCCCCCCCCCCCCCCCC

    SEQ KGKIELRIRSESNAQAIRLHEQDLNKRLEKELDVMTADHLREKNIMRADFNKTNELLEIN
25 SEG .....
PRD hhhhhhhhhhhcchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
COILS CCCCC.....

    SEQ AALQVSLEEEMEEKYLMRESKPEDIQMITE L KAMLTERDAQI IKKLIEDNKFYQLLVNRET
30 SEG .....
PRD hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
COILS .....

    SEQ NFNKVFNSSPTVGVINPLAKQKKKNDKSPTNR F VSPNL SALES GG VGNGHP NRLDP IPN
35 SEG .....
PRD hhhhhhhccccceeeehhhhhhhhhhcccccccccceecccccccccccccccccccccccccccccc
COILS .....

    SEQ SPVDHIEFNSSKPLPQPVP PKGPK TFLRYQ
40 SEG .....xxxxxxxxxxxxxx.....
PRD ccccccccceeeeeccccccccccccccccccccceeeeccc
45 COILS .....

```

55. group: testis derived .

DKFZphtes3_1bp3 encodes a novel 1bb3 amino acid protein without similarity to known proteins.

5 The novel protein is glutamine rich and contains a cell attachment RGD motif. According to the low number of ESTs and their origin the protein seems to be expressed ubiquitously at low levels.

No informative BLAST results; No predictive prosite, pfam or SCOP motif.

10 The new protein can find application in studying the expression profile of testis-specific genes.

15 putative protein

perhaps complete cds.

Sequenced by BMFZ

20 Locus: unknown

Insert length: 5411 bp -

25 Poly A stretch at pos. 5354, polyadenylation signal at pos. 5340

```

1  GGC GGCCAGG TGGAGGACCT GAGCAAGCAG CTCAAGCGTG TGGACGGCCA
51 GGTGCAGGGC ATCGCCACGC ACGTGCAGCA CTTCTCCAG GCCAGCGGGC
101 TTGACCTGGC CGCGCTAGAG TGGCCGGAGG AGCAGGAGGT GGGCGTGCGG
30 151 GCGTTCGATA GGGTGCGGAC TGGGAGTATC ATGAAGGACG CCGCCGAGGA
201 GCTCAGCTTT GCCAGGGTAC TTTTACAGCG GGTTGATGAA CTAGAGAAGC
251 TATTCAAAGA TCGGGAGCAA TTCCTGGAAC TAGTCAGCCG GAAGCTGAGT
301 TTGGTTCCTG GTGCAGAAGA AGTCACCATG GTCACCTGGG AAGAGCTGGA
35 351 GCAGGGCGATT ACGGACGGCT GGAGAGCCTC ACAAGCGGGC TCAGAAACAC
401 TTATGGGATT TTCTAAGCAC GGAGGGTTCA CTTCTTAAC ATCACCCTGAA
451 GGGACTCTAA GCGGAGACTC TACCAAGCAA CCAAGTATTG AGCAGGCTCT
501 GGATTCTGCC AGTGGTCTTG GCCCGGATCG GACTGCATCA GGATCTGGTG
551 GCACAGCACA CCCCTCTGAT GGGGTTTCCA GTAGGGAACA AAGCAAGGTC
601 CCCTCTGGTA CTGGGAGACA GCAGCAGCCG AGGGCCCGTG ATGAAGCTGG
40 651 CGTGCCACGA CTCCATCAGT CTTCTACATT CCAATTCAAA TCAGACTCAG
701 ATCGTCAACG GAGTAGAGAG AAGCTTACCT CGACACAACC AAGAAGAAAT
751 GCACGTCCTG GTCCAGTTCA ACAGGACTTA CCCTTGGCCA GAGACCAGCC
801 CAGTAGTG TG CCCGCTAGCC AGAGTCAGGT CCATCTAAGG CCAGATCGTC
851 GTGGGTTAGA ACCAACTGGC ATGAATCAGC CTGGATTAGT GCCTGCTAGC
45 901 ACTTACCCAC ATGGGTGTGGT ACCCCTCAGC ATGGGTGAGC TTGGTGTGCC
951 ACCACCTGAA ATGGATGATC GGGGAATTGAT ACCATTTGTC GTGGATGAGC
1001 AACGTATGTT GCCACCATCA GTACCTGGCA GAGACCAGCA AGGATTGGAA
1051 CTACCTAGCA CAGACCAACA TGGTCTGGTT TCAGTCAGTG CATATCAGCA
1101 TGGTATGACA TTTCTGGCA CAGACCAACG CAGTATGGAA CCACTTGGCA
50 1151 TGGATCAGCG TGGATGTGTA ATATCAGGCA TGGGTCAGCA AGGACTAGTA
1201 CCCCTGGTA TAGACCAGCA AGGATTGACA TTGCCTGTCT TCGATCAACA
1251 TGGCCTGGTT CTACCTTTTA CAGACCAGCA TGGTTTGGTA TCACCTGGTT
1301 TGATGCCAAT TAGTGACAGT CAGCAAGGTT TTGTGCAGCC CAGTTTGGAA
1351 GCAACTGGCT TCATACAACC TGGCACAGAG CAGCATGATT TAGTCCAGTC
55 1401 TGGCAGATTT CAGCGTGCTT TGGTGCAGCG TGGTGCATAT CAGCCTGGCT
1451 TGGTCCAACC TGGTGCAGAT CAGCGTGGTT TGGTCCGGCC TGGGAATGGAT
1501 CAGTCTGGTT TGGCCCAACC TGGTGCAGAT CAGCGTGGTT TGGTCTGGCC
1551 TGGGAATGGAT CAGTCTGGTT TGGCCCAACC TGGTAGAGAT CAGCATGGTT
```

	1601	TGATCCAGCC	TGGCACAGGT	CAGCATGATT	TGGTCCAATC	TGGCACAGGT
	1651	CAGGGTGTCT	TGGTACAGCC	TGGTGTAGAT	CAGCCTGGCA	TGGTCCAACC
	1701	TGGCAGATTT	CAGCGTGCTT	TGGTGCAGCC	TGGTGCATAT	CAGCCTGGCT
	1751	TGGTCCAACC	TGGTGCAGAT	CAGATTGATG	TGGTGC AACC	TGGTGCAGAT
5	1801	CAGCATGGTT	TGGTACAATC	TGGTGCAGAT	CAGAGTGATT	TGGTCAACC
	1851	TGGTGCAGTT	CAGCATGGTT	TGGTCCAACC	TGGAGTAGAT	CAGCGTGGTT
	1901	TGGCACAACC	TCGTGCAGAT	CATCAGCGTG	GTTTGGTCCC	ACCTGGTGCA
	1951	GATCAGCGTG	GTTTGGTCCA	ACCTGGTGCA	GATCAGCATG	GTTTGGTCCA
	2001	ACCTGGAGTG	GATCAGCATG	GTTTGGCACA	ACCTGGTGAA	GTTCAGCGTA
10	2051	GTTTGGTGCA	ACCTGGTATA	GTTTGGTCCA	GTTTGGTGCA	ACCTGGTGCA
	2101	GTTTGGTGCA	GTTTGGTGCA	ACCTGGTGCA	GTTTGGTGCA	GTTTGGTGCA
	2151	ACCTGGAGTG	GATCAGCGTG	GTTTGGTTCA	ACCTGGTGCA	GTTTGGTGCA
	2201	GTTTGGTCCA	ACCTGGTGCA	GTTTGGTGCA	GTTTGGTGCA	ACCTGGTGCA
	2251	GATCAGCGTG	GTTTGGTCCA	ACCTGGAGTG	GATCAGCGTG	GTTTGGTGCA
15	2301	ACCTGGAGTG	GATCAGCGTG	GTTTGGTCCA	ACCTGGAGTG	GATCAGCGTG
	2351	GTTTGGTCCA	ACCTGGTGCA	GATCAGCGTG	GTTTGGTGCA	GATCAGCGTG
	2401	GGTCAGCTGG	GTATGGTGCA	GCCTGGAATA	GGTCAGCAAG	GTATGGTGCA
	2451	ACCTCAGGCA	GATCCACATG	GCCTGGTACA	ACCTGGTGCC	TATCCTCTTG
	2501	GTTTGGTGCA	ACCTGGTGCA	TATTTGCATG	ATTTATCTCA	ATCTGGGACA
20	2551	TATCCACGTG	GTCTGGTGCA	GCCAGGAATG	GATCAGTATG	GTTTGGTGCA
	2601	ACCTGGTGCA	TATCAGCCAG	GCTTGATAGC	ACCAGGCACA	AAGCTTCGTG
	2651	GCTCTTCAAC	ATTCCAGGCA	GATTCTACAG	GTTTATATATC	AGTACGTCCA
	2701	TATCAACATG	GTATGGTACC	TCCTGGCAGA	GAACAATACG	GCCAGGTGTC
	2751	ACCACTCCTA	GCCAGTCAAG	GTTTGGCATC	ACCTGGTATA	GATCGAAGGA
25	2801	GTTTGGTACC	ACCAGAACT	TATCAGCAAG	GTTTGGTGCA	TCCTGGCACA
	2851	GACCAGCACA	GCCCAATACC	ACTGAGTACA	GGTTTGGGAT	CTACACACCC
	2901	AGATCAACAG	CATGTGGCAT	CACCTGGCCC	AGGTGAGCAT	GACCAGGTAT
	2951	ACCCAGATGC	AGCTCAGCAT	GGCCATGCTT	TCTCTCTCTT	TGACAGTCAT
	3001	GATTCAATGT	ATCCTGGTTA	TCGTGGCCCA	GGGTATCTAA	GTGCTGATCA
30	3051	GATGGGCTAG	GAAGGTTTGG	ATCCAAATAG	AACACGAGCC	TCGGACCGAC
	3101	ATGGAATTCC	TGCCCAGAAAG	GCCCCAGGCC	AAGATGTAC	TCTTTTACAGG
	3151	AGTCCAGACT	CCGTGACCG	AGTCTTATCA	GAAGGGAGCG	AAGTCTCGAG
	3201	TGAAGTCCTG	AGTGAGCGAC	GCAATTCACT	GCGTAGAATG	AGTTCTAGTT
	3251	TCCCCACGGC	AGTGGAGACA	TTTCATCTGA	TGGGAGAGCT	CAGTAGCCTC
35	3301	TATGTGGGGC	TAAAGGAGAG	TATGAAGGAT	CTGGATGAGG	AGCAGGCCGG
	3351	CCAAACCGAC	TTGGAGAAGA	TCCAGTTCCT	GCTGGCACAG	ATGGTCAAAA
	3401	GGACCATAAC	TCCTGAACTG	CAGGAGCAGC	TGAAGACCGT	AAAGACGCTA
	3451	GCCAAAGAAAG	TTTGGCAGGA	GAAAGCAAAA	GTGGAAAGGC	TGCAGAGGAT
	3501	CCTGGAAGGG	GAAGGGAATC	AAGAAGCAGG	GAAGGAACTG	AAAGCTGGAG
40	3551	AGCTGAGATT	GCAGCTGGGT	GTCCTCAGAG	TCACCGTGCG	TGACATAGAA
	3601	AAGGAGCTGG	CCGAGTTGAG	GGAGAGCCAA	GACAGGGGCA	AGGCTGCCAT
	3651	GGAAAATTCT	GTCTCTGAAG	CCTCCCTTTA	CCTGCAGGAC	CAGTTGGACA
	3701	AGCTCAGGAT	GATCATTGAG	AGCATGCTGA	CCTCCTCCTC	CACGCTCCTG
	3751	TCCATGAGCA	TGGCCCCGCA	CAAGGCCAC	ACCTTGGCTC	CTGGCCAGAT
45	3801	CGACCCTGAG	GCCACCTGTC	CAGCCTGCAG	CCTGGATGTG	AGCCATCAGG
	3851	TCAGCACGCT	GGTGCGGCGC	TATGAGCAAC	TCCAAGACAT	GGTCAACAGC
	3901	CTGGCCGTCT	CCCGACCCTC	CAAGAAGGCC	AAGCTCCAGA	GACAGGACGA
	3951	GGAGCTGCTG	GGCCGTGTGC	AGAGTGCCAT	CCTGCAGGTG	CAGGGTGACT
	4001	GCGAGAAGCT	CAACATCACC	ACCAGCAACC	TCATCGAGGA	CCATCGGCAG
50	4051	AAACAGAAGG	ACATTGCTAT	GCTGTACCAG	GGTCTGGAGA	AGCTCGAAAA
	4101	GGAAAAGGCC	AACAGGGAGC	ACCTGGAGAT	GGAGATCGAT	GTGAAAGCCG
	4151	ACAAGAGTGC	TCTGGCCACC	AAAGTGAGCC	GTGTCCAGTT	TGATGCCACC
	4201	ACGGAGCAGC	TGAACCACAT	GATGCAGGAG	CTGGTGGCCA	AGATGAGCGG
	4251	GCAGGAGCAG	GACTGGCAGA	AGATGCTGGA	CAGGCTGCTC	ACAGAGATGG
55	4301	ACAACAAGCT	GGACCGCCTG	GAGCTGGACC	CAGTGAAGCA	GTTGCTGGAG
	4351	GATCGGTGGA	AATCGGTGCG	ACAGCAGCTC	AGGGAGCGCC	CCCCACTCTA
	4401	CCAGGCAGAC	GAGGCGGCTG	CCATGCGGAG	GCAGCTCCTG	GCACATTTCC
	4451	ACTGCCTCTC	ATGTGACCGG	CCCTTGGAGA	CACCTGTGAC	TGGACATGCC

```

4501 ATCCCCGTGA CCCCCGCGGG TCCAGGCCTA CCTGGGCACC ATTCCATCCG
4551 CCCCCTACACG GTGTTTGAAC TGGAGCAGGT CCGGCAGCAT AGCCGCAACC
4601 TCAAGCTGGG CAGCGCCTTC CCTCGGGGTG ACCTGGCGCA GATGGAGCAG
4651 AGCGTGGGGC GCCTGCGCTC CATGCACTCC AAGATGCTGA TGAACATTGA
5 4701 GAAGGTGCAG ATCCACTTCG GGGGCTCCAC CAAGGCCAGC AGCCAGATAA
4751 TCCGCGAGCT GCTGCACGCC CAGTGCCTGG GCTCCCCCTG CTACAAACGG
4801 GTGACAGATA TGGCTGATTA CACCTACTCA ACTGTGCCCC GCGCTGCGG
4851 GGGCAGCCAC ACCCTCACCT ACCCTACCA CCGCAGCCGC CCGCAGCACC
4901 TTCCCCGGGG CCTGTATCCT ACTGAAGAGA TCCAGATTGC CATGAAGCAT
10 4951 GATGAGGTGG ACATCTTGGG CCTGGATGGC CACATTTACA AGGGACGGAT
5001 GGACACAAGG CTGCCAGGCA TCCTCCGAAA AGACAGCTCA GGGACCTCAA
5051 AGCGCAAGTC CCAGCAGCCC AGGCCCCACG TGCACAGGCC GCCATCCCTC
5101 AGCAGCAATG GCCAGCTGCC CTCTCGGCCA CAGAGCGCCC AGATTTCCGC
5151 TGGCAACACC TCAGAAAGAT AGACCTTCCT CCGAGGGCCG TCTCTCCAG
15 5201 CCGAACACAG CCCACCCGCC CAGCTCCGCC TCGGTGGCAA ACAGGGGGCT
5251 GGAGAGGCAC GTGGACATGC CTCCTGGGGA GGGGCTCGAG GAGCCCACGC
5301 GGGGGCCGCG GTCCAGCACC GCTCAGTGAG CGGAGGTGTA AATAAACATT
5351 CAGGAGGAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
5401 AAAAAAAAAA A

```

BLAST Results

25 No BLAST result

Medline entries

30 No Medline entry

Peptide information for frame 1

ORF from 181 bp to 5169 bp; peptide length: 1663
 Category: putative protein
 40 Classification: no clue
 Prosite motifs: RGD (1482-1484)

```

1 MKDAAEELSF ARVLLQRVDE LEKLFKDREQ FLELVSRKLS LVPGAEEVTM
45 51 VTWEELEQAI TDGWRASQAG SETLMGFSKH GGFTSLTSPE GTLSGDSTKQ
101 PSIEQALDSA SGLGPDR TAS GSGGTAHPSD GVSSREQSKV PSGTGRQQQP
151 RARDEAGVPR LHQSSTFQFK SDSDRHRSRE KLTSTQPRRN ARPGPVQQDL
201 PLARDQPSV PASQSQVHLR PDRRGLEPTG MNQPGLVPA S TYPHGVVPLS
251 MGQLGVPPPE MDDRELIPFV VDEQRMLPPS VPGRDQQGLE LPSTDQHGLV
50 301 SVSAYQHGMT FPGTDQRSME PLGMDQRCV ISGMGQQLV PPGLDQQGLT
351 LPVVQDHGLV LPFTDQHGLV SPGLMPISAD QQGFVQPSLE ATGFIQPGTE
401 QHDLIQSGRF QRALVQRGAY QPGLVQPGAD QRGLVRPGMD QSGLAQPGAD
451 QRGLVWPGMD QSGLAQPGRD QHGLIQPGTG QHDLVQSGTG QGVLVQPGVD
501 QPGMVQPGRF QRALVQPGAY QPGLVQPGAD QIDVVQPGAD QHGLVQSGAD
55 551 QSDLAQPGAV QHGLVQPGVD QRGLAQPRAD HQRGLVPPGA DQRGLVQPGA
601 DQHGLVQPGV DQHGLAQPGE VQRSLVQPGI VQRGLVQPGA VQRGLVQPGA
651 VQRGLVQPGV DQRGLVQPGA VQRGLVQPGA VQHGLVQPGA DQRGLVQPGV
701 DQRGLVQPGV DQRGLVQPGM DQRGLIQPGA DQPGLVQPGA GQLGMVQPGI

```



```

5  751 GQQGMVQPGA DPHGLVQPGA YPLGLVQPGA YLHDLSSQSGT YPRGLVQPGM
801 DQYGLRQPGA YQPGLIAPGT KLRGSSTFQA DSTGFISVRP YQHGMPVPPGR
851 EQYGVQVSPLL ASQGLASPGI DRRSLVPPET YQQGLMHPGT DQHSPIPLST
901 GLGSTHPDQD HVASPGPGEH DQVYPDAAQH GHAFSLFDSD DSMYPGYRGP
10 951 GYLSADQHGQ EGLDPNRTRA SDRHGIPAQK APGQDVTLFR SPDSVDRVLS
1001 EGSEVSSEVL SERRNSLRMM SSSFPTAVET FHLMGELSSL YVGLKESMKD
1051 LDEEQAGQTD LEKIQFLLAQ MVKRTIPPEL QEQLKTVKTL AKEVWQEKAK
1101 VERLQRILEG EGNQEAGKEL KAGELRLQLG VLRVTVADIE KELAELESQ
1151 DRGKAAMENS VSEASLYLQD QLDKLRMIIE SMLTSSSTLL SMSMAPHKAH
10 1201 TLAPGQIDPE ATCPACSLDV SHQVSTLVRR YEQLQDMVNS LAVSRPSKKA
1251 KLQRQDEELL GRVQSAIQLV QGDCEKLNIT TSNLIEDHRQ KQKDIAMLYQ
1301 GLEKLEKEKA NREHLEMEID VKADKSALAT KVSrvQFDAT TEQLNHMMQE
1351 LVAKMSGQEQ DWQKMLDRL TEMDNKLDRL ELDPVKQLLE DRWKSRLRQL
1401 RERPPLYQAD EAAAMRRQLL AHFHCLSCDR PLETPVTGHA IPVTPAGPGL
15 1451 PGHHSIRPYT VFELEQVRQH SRNLKLGSAF PRGDLAQMEQ SVGRLRSMHS
1501 KMLMNIKQVQ IHFGGSTKAS SQIIRELLHA QCLGSPCYKR VTDMAQYQYS
1551 TVPRRCGGSH TLTPYHRSR PQHLPRGLYP TEEIQIAMKH DEVDILGLDG
1601 HIYKGRMDTR LPGILRKDSS GTSKRKSQQP RPHVHRPPSL SSNGQLPSRP
1651 QSAQISAGNT SER
20

```

BLASTP hits

25 No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_16p3, frame 1

No Alert BLASTP hits found

30

Pedant information for DKFZphtes3_16p3, frame 1

Report for DKFZphtes3_16p3.1

35

```

[LENGTH] 1723
[MW] 187354.98
[pI] 6.19
40 [HOMOL] TREMBL:AF025461_4 gene: "M01D1.5"; Caenorhabditis
elegans cosmid M01D1. 1e-47
[FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae,
YDL058w] 8e-07
45 [FUNCAT] 08.07 vesicular transport (golgi network, etc.) [S.
cerevisiae, YDL058w] 8e-07
[FUNCAT] 99 unclassified proteins [S. cerevisiae, Y0R216c]
2e-04
[FUNCAT] 11.04 dna repair (direct repair, base excision repair
and nucleotide excision repair) [S. cerevisiae, YKR095w] 0.001
50 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YKR095w]
0.001
[BLOCKS] PR01098C
[BLOCKS] BP02308D
[BLOCKS] PR00543H
55 [BLOCKS] PR00210G
[BLOCKS] PR00210E
[BLOCKS] BP04236A
[PIRKW] RNA binding 3e-06

```

[illegible]

SEQ TYPHGVVPLSMGQLGVPPPEMDDRELIPFVVDEQRMLPPSVPGRDQQGLELPSTDQHGLV
SEG
PRD CC
COILS
5
SEQ SVSAYQHGMTFPGTDQRSMEPLGMDQRCVISGMGQGLVPPGIDQGLTLPVVDQHGLV
SEG
PRD CC
COILS
10
SEQ LPFTDQHGLVSPGLMPISADQGGFVQPSLEATGFIQPGTEQHDLIQSGRFQRALVQRGAY
SEG
PRD CC
COILS
15
SEQ QPGLVQPGADQRLVVRPGMDQSGLAQPGADQRLVWPGMDQSGLAQPGRDQHGLIQPGTG
SEG
PRD CC
COILS
20
SEQ QHDLVQSGTGQGVLVQPGVDQPGMVQPGRFQRALVQPGAYQPGLVQPGADQIDVVQPGAD
SEG
PRD CC
COILS
25
SEQ QHGLVQSGADQSDLAQPGAVQHGLVQPGVDQRLAQPRADHQRLVPPGADQRLVQPGA
SEG
PRD CC
COILS
30
SEQ DQHGLVQPGVDQHGLAQPGEVQPSLVQPGIVQRLVQPGAVQRLVQPGAVQRLVQPGV
SEG
PRD CC
COILS
35
SEQ DQRLVQPGAVQRLVQPGAVQHGLVQPGADQRLVQPGVDQRLVQPGVDQRLVQPGM
SEG
PRD CC
COILS
40
SEQ DQRLIQPGADQPGLVQPGAGQLGMVQPGIGQGMVQPGADPHGLVQPGAYPLGLVQPGA
SEG
PRD CC
COILS
45
SEQ YLHDLQSGTYPRGLVQPGMDQYGLRQPGAYQPGLIAPGTKLRGSSTFQADSTGFISVRP
SEG
PRD CC
COILS
50
55

[illegible][illegible]

```

      SEQ    QEQLKTVKTLAKEVWQEAKAKVERLQRILEGEGNQEAGKELKAGELRLQLGVLRVTVADIE
      SEG     .....
30 PRD       hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
   COILS     .....CCCCCCCCCCCCCCCCCCCCCCCCCC

```

[illegible]

```

40  SEQ  TLAPGQIDPEATCPACSLDVSHQVSTLVRRYEQQLQDMVNSLAVSRPSKKAKLQRQDEELL
    SEG  .....
    PRD  hhccccccccccccccccchhhhhhhhhhhhhhhhhhhhhhhccccchhhhhhhhhhhhhh
    COILS

```

[illegible]

```

50 .....  

    SEQ VKADKSALATKVS RVQFDATTEQLNHMMQELVAKMSGQE QDWQKMLDRLLTEM DNKL DRL  

    SEG .....  

    PRD hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhhhhhhh  

55 COILS .....

```

SEQ ELDPVKQLLEDRLKSLRQQLRERPPLYQADEAAAMRRQLLAHFHCLSCDRPLETPVTGHA

5

10

15

20

25

Prosites for DKFZphtes3_16p3.1

35

-324-

DKFZphtes3_17i21

5 group: transmembrane protein

DKFZphtes3_17i21 encodes a novel 224 amino acid protein without similarity to known proteins.

10 The novel protein contains 2 transmembrane regions. ESTs can be found in testis, retina and brain.
No informative BLAST results; No predictive prosite, pfam or SCOP motife.

15 The new protein can find application in studying the expression profile of testis-specific genes and as a new marker for testicular cells.

20 unknown protein

Pedant: contains signal peptide(frame 1) and TRANSMEMBRANE 2 (frame 2)

25 perhaps complete cds.

Sequenced by GBF

Locus: unknown

30

Insert length: 1518 bp

Poly A stretch at pos. 1480, polyadenylation signal at pos. 1454

```

35      1 GCCAGACAGC TAGGTGTCAT TCAGGGCTGG TGTCTCTGT CCAGGCCATC
      51 ATGGCCTCCA CTGCCGGCTA CATCGTCTCC ACCTCCTGCA AGCACATCAT
     101 TGATGACCAA CACTGGCTGT CCTCTGCCTA CACGCAATTT GCTGTGCCCT
     151 ACTTCATCTA CGACATCTAC GCCATGTTCC TCTGTCACTG GCACAAGCAC
     201 CAGGTCAAAG GGCATGGAGG GGACGACGGA GCGGCCAGAG CCGCGGGCAG
     40  251 CACGTGGGCC ATAGCGCGTG GCTACCTGCA CAAGGAGTTC CTCATGGTGC
     301 TCCACCATGC CGCCATGGTG CTGGTGTGCT TCCCACTCTC AGTGGTGTGG
     351 CGACAGGGTA AGGGAGACTT CTTTCTGGGT TGCATGTTGA TGGCAGAGGT
     401 CAGCACGCCC TTCGTCTGCC TTGGCAAGAT CCTCATCCAG TACAAGCAGC
     451 AGCACACACT GCTGCACAAG GTGAACGGGG CCCTGATGCT GCTCAGCTTC
     45  501 CTCTGCTGCC GGGTGTGCT CTTTCCCTAC CTGTACTGGG CCTACGGGCG
     551 CCATGCCGGC CTGCCCCCTGC TGGCCGTGCC CCTGGCCATC CCTGCCACG
     601 TCAACCTGGG CGCTGCGCTG CTCCTGGCCC CTCAGCTCTA CTGGTTCTTC
     651 CTCATCTGCC GTGGGGCCTG CCGCCTCTTC TGGCCCCGCT CCGGGCCGCC
     701 CCGGGCCTGC CAGGGCCAGG ACTGAGGCCG GGGGCCGGGA CCCTCCCCCT
     50  751 CCCCACCCCC ACCCCCGTGG AGACAGGGCT CTGGGGCTGA TGGCTGGGGT
     801 TGGGAGCCAG GGTCTCTTTG CCCGGACAAC CCCAGGACTG ACGATGACCC
     851 CGAAAGGGAA GAGGCCCAT CTCTCGGGGA CTGAGGGGGT GGAGAGAGGG
     901 GACCTCTTCC CCCTACTCTG CCCCCTTCCT GCACACCTT GCGCTGGAGG
     951 AGGGGAGGGG GCACGCCTC CCACCCACTG AGGGCAGGAG GGCTTGTGGG
    55 1001 GAGGGACACC AACAGGGTTT CAAGGGGACC AGGAGTCAGA ATGTGGGGAG
    1051 ACGCCTCTGC CAAGGCCATC CCAGCCCCTA TGCTGCCATC CCCCAGGGCT
    1101 CCCCATCACC CGAGAGGAGA GGACGCCCCA ACTAACCCCC GCTGGCCCTC
    1151 GGGCCTCCCC AGTGGCCGGC TGCAACCACG GCTCCTCTCC AGGGTAGGCC

```

1201 AGCTTGAGGA ATCTTATTTA TTTTATTTAT TTACCCAAAT TTGAACTAGT
 1251 CTGTTGGGTT GGGGGAAGGA GGTGGCTGCT ACCCCCAAGC CTTCCCAGTG
 1301 CTGACAACCC CGGGGGCAGG CGAGGGCGCC CAGTCCCTCA CCATCGGCTG
 1351 CACATCGCGC CCTCGGGCCC TGCCATGTCC CTGGTGCTAC TGACCTCTCA
 5 1401 AGGCTTCCTC CAATCTGGGG TCGGGGGACC CTGGGAGGTG CTTTACAGAC
 1451 CGCTAATAAA AGACGATCTG CGTGAACGCC AAAAAAAAAA AAAAAAAAAA
 1501 AAAAAAAAAA AAAAAAAAAA

10

BLAST Results

No BLAST result

15

Medline entries

No Medline entry

20

Peptide information for frame 3

25

ORF from 51 bp to 722 bp; peptide length: 224

Category: putative protein

Classification: Transmembrane proteins unclassified

30

1 MASTAGYIVS TSCKHIIDDQ HWLSSAYTQF AVPYFIYDIY AMFLCHWHKH
 51 QVKGHGDDG AARAPGSTWA IARGYLHKEF LMVLHHAAMV LVCFPLSVVW
 101 RQGGKDDFFLG CMLMAEVSTP FVCLGKILIQ YKQQTLLHK VNGALMLLSF
 151 LCCRVLLFPY LYWAYGRHAG LPLLAVPLAI PAHVNLGAAL LLAPQLYWFF
 201 LICRGACRLF WPRSRPPAC QAQD

35

BLASTP hits

40 No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_17i21, frame 3

No Alert BLASTP hits found

45

Pedant information for DKFZphtes3_17i21, frame 3

Report for DKFZphtes3_17i21.3

50

[LENGTH] 224

[MW] 25224.11

[pI] 9.03

55

[HOMOL] TREMBLNEW:AF181646_1 gene: "BcDNA.GH1232b";
 product: "BcDNA.GH1232b"; Drosophila melanogaster BcDNA.GH02340
 (BcDNA.GH02340) mRNA, complete cds. 9e-20
 [BLOCKS] PRO0632H

[[BLOCKS]] PRO0904A
[[BLOCKS]] BLO1243C
[[KW]] TRANSMEMBRANE 2
[[KW]] LOW_COMPLEXITY 6.25 %

5

SEQ MASTAGYIVSTSCKHIIDDQHWLSSAYTQFAVPYFIYDIYAMFLCHWHKHQVKGHGGDDG
SEG
PRD cccccccccccccccccccchhhhhhhhhhhheeehhhhhhhhhhhhhhhhhhhhcccccccccc
MEM

10

SEQ AARAPGSTWAIARGYLHKEFLMVLHHAAMVLVCFPLSVVWRQGKGDFFLGCMLMAEVSTP
SEG
PRD cccccccccccccccchhhhhhhhhhhhhhhhhhhhhccccccccccccchhhhhhhhhhhccc
MEMMMMMMMMMMMMMMMMMMM.....

15

SEQ FVCLGKILIQYKQQHTLLHKVNGALMLLSFLCCRVLLFPYLYWAYGRHAGLPLLAVPLAI
SEGxxxxxxxxxxxx
PRD ccchhhhhhhhhhhhhhhhhccchhhhhhhhhhhheeeccceeecccccccccccccccccc
MEMMMMMMMMMMMMMMMMMMM.....

20

SEQ PAHVNLGAALLLAPQLYWFFLICRGACRLFWRPRSRPPACQAD
SEG xx.....
PRD cchhhhhhhhhhhccceeeeeecccccccccccccccccccccccccc
MEM

25

(No Prosite data available for DKFZphtes3_17i21.3)

30 (No Pfam data available for DKFZphtes3_17i21.3)

DKFZphtes3_18n14

5 group: transcription factors

DKFZphtes3_18n14 encodes a novel 377 amino acid protein with similarity to human giantin.

10 Giantin is discussed as an autoantigen in rheumatoid arthritis. The novel protein contains a leucine zipper and a putative Helix-loop-helix DNA-binding domain. Therefore it might be a novel transkription factor. Most EST hits are from testis and germ cells.

15 The new protein can find application in modulation of gene expression and in expression profiling.

20 unknown protein

see DKFZphtes3_30i23
wrong orientation
perhaps complete cds.

25 Sequenced by MediGenomix

Locus: /chromosome="16"

30 Insert length: 5282 bp
Poly A stretch at pos. 5242, polyadenylation signal at pos. 5227

```

      1 CCGGCACCCG GAGCTCCTGG GCACACGGCA TTGGCAGGGG CCGCTTCGGC
35    51 AGAGTGATGA CTGATGATGA GTCCGAGAGC GTCCTCTCCG ACTCCCATGA
      101 AGGGTCGGAG CTGGAGCTGC CTGTTATCCA GCTGTGCGGG CTGGTGGAGG
      151 AGCTCAGCTA TGTAACCTCT GCTCTCAAAA CTGAGACTGA GATGTTTGAG
      201 AAATATTACG CTAACTGGA GCCCAGGGAT CAGCGACCTC CACGATTATC
      251 AGAAATTAAA ATATCAGCAG CAGATTATGC ACAGTTTCGA GGCAGGCGTA
40    301 GATCCAAATC CCGGACAGGT ATGGACCGTG GGGTAGGCCT GACTGCCGAC
      351 CAAAAACTTG AGCTGGTACA AAAAGAGGTT GCGGACATGA AGGATGACTT
      401 ACGACACACA AGGGCAAATG CGGAACGCGA CCTGCAGCAT CACGAGGCGA
      451 TCATTGAGGA GGCTGAAATT CGATGGAGTG AAGTTTCGAG AGAAGTGCAT
      501 GAGTTTGAAA AAGATATTCT AAAAGCCATA TCCAAGAAGA AAGGGAGTAT
45    551 TTTGGCCACT CAGAAAGTGA TGAAATACAT TGAGGACATG AACCGCCGGA
      601 GGGATAATAT GAAGGAGAAA TTACGTTTGA AAAATGTTTC TCTCAAAGTT
      651 CAGAGGAAAA AAATGCTTTT ACAATTGAGG CAGAAGGAAG AGGTGAGTGA
      701 GGCCCTTCAC GATGTTGATT TTCAGCAGTT GAAGATAGAG AACGCTCAAT
      751 TTCTTGAGAC AATTGAAGCA AGGAATCAAG AACTGACCCA GCTAAAGCTG
50    801 TCATCTGGAA ACACTCTGCA GGTTCCTCAAT GCCTACAAAA GCAAGCTTCA
      851 CAAGGCAATG GAAATATACC TCAATCTGGA CAAGGAGATC TTGCTGAGAA
      901 AAGAGCTACT TGAAAAAATT GAAAAAGAAA CACTACAAGT AGAGGAGGAC
      951 CGGGCCAAAG CCGAGGCAGT GAATAAGAGG CTCCGGAAGC AGCTGGCCGA
55   1001 GTTCCGGGCA CCACAGGTGA TGACTTACGT CCGGGAGAAG ATCTTAAATG
      1051 CGGACCTGGA GAAGAGCATC AGGATGTGGG AAAGGAAAGT GGAGATAGCA
      1101 GAGATGTCTT TAAAAGGCCA TCGTAAGGCT TGGGAATCGAA TGAAAATAAC
      1151 CAATGAGCAG TTGCAGGCAG ATTACCTTGC TGGGAAGTAG CCAGAGGCAG
      1201 GCCACGGCTT ACAGACCACT ACATGACCTA TAAAAGTAAT CAGCTCCTTT

```

	1251	CTAGTCACGG	GCTCCTCTCA	CTGTTCCCTG	TCTGCCTGGT	GTTCCTCAACC
	1301	CCCCACCCAG	GCTGAGTATC	ATCTCCTGGG	CCACATCTGC	CCATGGGGAG
	1351	TGTTTTTACA	GCCTGGCCCC	TGGAACTGTT	ACCACTGAAA	GAACCACAGG
5	1401	GCACTCTAAT	GGTTTGACAC	TTGTTAGCCA	GCATTTAGTT	CACAAGCATA
	1451	GTGAAAGTGA	CCTTCCCACA	CCTGGGAGAG	GGATAGAGGA	GGGAGAGCCA
	1501	GCCCAGTGTA	TGCCATGGGC	TTATCCGTGG	CAGCCCCAGT	GTGCAACTAT
	1551	CAAAAACAGA	CATCAAAACA	GCATGGTGAA	TGCCTGGCAC	TCAGCATTCT
	1601	CAGTTTACTC	TTCAGTTTGG	TGGGGTAGCT	CCTGGACTAG	ATACTGCTGC
	1651	AAAAGAAAAC	AAGCACGAAG	GAAACCAAGA	TGATTTCTTC	GGGCTGATAC
10	1701	AACCTGTTCT	GACCTGCAAA	AATCCTACCT	TCCCCACCT	CCCCACCGTA
	1751	ATAGTCATAG	TATAAGGGTT	GTACAGACGC	CTCAGGAGAC	CTGCCTGATT
	1801	CCTTTACATC	CTTCTCCCTA	ACATCTAGAC	TATCTCTAGA	GCTGTTTCTT
	1851	AGTCGTGAAT	GCGTGATGGT	CCTTCTTTGT	CCCTGCAAGT	ATGATCCAAC
	1901	ATGGCCCAAGT	TCAGAATCAG	AATATGTCTT	CTGTGTCATG	GTGGCATTTG
15	1951	GTCCATGGTG	GGAGAAAGAA	ATCAACTTTT	CCCAGTGGTG	GAGTGAGGAC
	2001	AGGGGAGGGC	CGGCCCTCTC	AGCCTTGGAT	GTGATCCATT	TGCTGTAGTC
	2051	TTCCACCTTG	GTGTACAGAA	ACAGGCCAGG	GCACGTCTCA	CCACCGAAGT
	2101	TCAGGACTCC	TCTCAGAACC	CACAGATCGA	ACTGCTGTAG	CTGGCACATC
	2151	ATTGGGCTTC	CTGGGTCCCC	CTGTGATAAA	AGACAGAAGG	CTTCAAGTCT
20	2201	TAGAAAAACT	AGTTTTTGT	GTAAATCTAT	CCTTGTGCAA	TATACTGTTT
	2251	GTTCTAGAAA	TGTTTTACGC	TGGTTCTCAC	TGGAAATGGG	GCAAATTATA
	2301	GGATACAATT	TCAAATCTAG	GCAGCCACCA	CCACAAATTC	CAACAAGATG
	2351	ACTTTTCTTT	TTATTATGCA	AATTAGCTGT	GGACTTCTGC	TGATTGCCTA
	2401	TAGCTTCTTG	GTTTCATATTT	CATTTTCTTG	CCCCTTTCCA	GTCTTTTGGC
25	2451	CAAACCTTCC	CTCTCTTCTG	GCTTCTCATT	CCTGAAATGT	TGGTGTTTGT
	2501	TTCTGTTTTG	TCCTGAAATG	CTCACATTTT	CCCTTCTCTG	CCTTGCTTCA
	2551	ACCCTTAGTG	TAAGCCACTT	CCTGCCACCT	GGCAACTGCT	TACCAGCCTG
	2601	GCTGGCCGTG	CTCTGGGTCT	TCCCTACTCC	CAATGGAGCA	GTCTCTTGGG
	2651	ACTTGGGAAT	TCTGCCACAT	ACACTTTATC	TAACTTAAAG	TGACGGAGTA
30	2701	GAAGCTTGCC	ATCATTAGCT	AGATATGGGA	CCCTGGCAAG	TGACCAAATC
	2751	CTCTCTGAGC	CAAGGTGGGA	ACACAGTTAA	TGCCCTGTAAC	ACGTGCTGAG
	2801	CACAGCACAG	TGCCCTGGCAC	ACAGCAAACA	CTCAATAGAA	TATTAGCTAC
	2851	CATCATCCTG	ATGTCGCTAT	AAAGGCCAGC	ATTTTTCTGA	AAAGTTGGGG
	2901	AAAATGGGAA	AAGCAACAAG	GCAACTAGTA	GGTATCACTT	ACCTTACCTG
35	2951	CCCAGACCCC	ACACCCCTAG	GTCTCCTCTC	AAAGGAATTC	CTGCCCCCTC
	3001	CATGGCCCAT	CTTGGTCCGA	GAAGGGGGTG	GTCAATCCCCA	GGCTAGCCAG
	3051	CCACTTCTGA	CCTGTGTGGC	CTGCCTGGCT	GGAAAGGCCCA	GGCAATGACA
	3101	TGTTGCTCTC	GCAGTTTGGG	CTGAGACATG	GAATGGGGCC	GCAATTAAACA
	3151	ACAGGAAACA	ATCTGAACAG	ACTGAACCAC	GAGCAGCAGA	AAGGCAGAAG
40	3201	AGCAGCCGCT	TCAGCCCTTT	ACCATCCGAG	ACCTGGGTGT	GTGGTCTGTC
	3251	TTGGTCACTC	TCTCTGTCTC	TCTTTCTCTC	TTTCTTTCTC	TGTCCCCAAG
	3301	GCTGGAGTGC	AGTGGTGCAA	TCTTGGCTCA	CTGCAACCTC	CACCTCTGGG
	3351	ATTCAAGCAA	TTCTCCCACC	TCAGCCTCTC	GAGTAGCTGG	GGCTACAGCT
	3401	ATGCGCCACC	ATGCCCAGCT	AATTTTTTTT	TTTTTTTTTT	GAGATGGAGT
45	3451	CTTGCTCTGT	CCCCCATGCT	GGAGTGCACT	GGCATGATCT	CGGCTCGCTG
	3501	CAACCTCCCT	CTCCTGGGTT	CAAGCGATTC	TCCTACCTCA	GCCTCCCCAG
	3551	TAGTGGGAT	TACAGGCCCC	CACCACCACA	CCTGGCTAAT	TTTTATTTTT
	3601	AGTAGAGATG	GGGTTTCAAC	ATGTTGGCCA	GGCTGGTCTC	GAACTCTTGA
	3651	CCTCATGATC	CACCCGCCTC	GGCCTCCCCA	AGTGTTGGGA	TTACAGGCGT
50	3701	GAGCCACTGC	ACCCGGCCTA	ATTTCTGTAT	TTTTAGTAGA	GATGGGGTTT
	3751	CACGATGTTG	GCCAGGCTGG	TCTTAATCTA	ACTTCAAGTG	ATCTGCCCCG
	3801	CTCGCCCTCT	CAAAGTGCTG	GGATTAGGCA	TGAACCTACCA	TGCCCAGTGG
	3851	GGTATTCTCT	TTCAATAAAG	CTCCTCTTTT	CCAAGGAAGC	CACACCAGAA
	3901	CAGAGATGAA	GACCAGTGGG	AAAACATGGG	AGCAACTCCG	TGGGCAGGCC
55	3951	AGCGGGGAGG	CCATGCTGCA	AAGCTGCCGT	GATTCCCTGG	TGATCTCTCA
	4001	GCAGGCCAAG	GCCAGACATG	TGAGGAAGGC	CTTGAGGACT	CTATTCTGTG
	4051	CCTCTCCTTG	GATGGAAGGG	GGTGCTTTAG	TGTGGCACTC	CTGACTTTTC
	4101	AATTGACTGG	TGAAGAGGCC	CTTGTGTGCA	CCTCACTATG	TCTGCCTAGG

```

4151 TCATGGGGGC TCCCTGGCCA AGAATGACGT GGTCCCCCT TTCATCAGTC
4201 CGATTGCGAG TTTGTCTTAA CTGTAGTGGT ATAGCCAGAG CAAGAAAAAG
4251 AATGTGATTT AGGACAAATG ATTGGATGAG TGATTGGTAG ATGTCCTCAG
4301 CTATGGCGTG GTTTTGCAAG TCACTGTTCC ACCCACCTGG GCACAGCATA
5 4351 TACGCTTTTT CTCTTCCCCA TAATCCCCTA GGGGCTGCGA CTTCTGAAGC
4401 ACAAGAGGCA GAGGCGAACA GCTCCAGGTG CCCCTCTGGA GCTACCTTAC
4451 CTCATCTCCC AAGGGAGCGG CCACAGCCCA GAGTGGGGTC TTTCATTTTG
4501 TGATCTTTTC CCTTGACATT CAGCAAAAAGC CCTGACAGTG GTAGAATAAA
4551 GGCAGGATGG GTGAGTGCAG AGTGATTCTG CTTTGTGTTGG GTTTCAGGGA
10 4601 AACCACATAGG CAGATTCTGA ACCTGGTGGT TGATTCTACA TGTGGGAATT
4651 GTGGCTTTGA AGACCTCTGG ACATGAGAAC ATATTTCCAA GACAGAGGAT
4701 TCTATGGGGA CGGGTCACCA TTAATGGTG TGCAAGCATA ATTCTGTTCA
4751 AAAATGAAGG CATGTTTAGA GGTGTGTAC AGTTAAAAAC CAACCTGAAC
4801 TTTGCAGTTA GATTTTAAAA GATGGTCAGT TAGAGTAGAA ATAGCTTAGA
15 4851 ATATTCCATT GAGTCTAAGA TACAGTTAGA AATCAACATC TTTGAAATTA
4901 GGGTGTGTCT TTTAATCAGT TGATGTCAGA GTTTAACGGG CAGCATTTTT
4951 TTCTTTCTTG GGATTACAAA AAATGATGGT GCATTCTATA ATTGGCAGCA
5001 TCTTAGATCT GAGGAAGTAT GATACTTGT TGACGGAATG GTTGACGGCA
5051 GAATTTTGTT AAAAAGCTAT ATCTTCACTG TATTTTAAACA CATTATCTAA
20 5101 TTTAAGAAAT GTTAAAGATC CCCCACCTGG CAGAGGACCC AGTACAAAAT
5151 AGGCACTCAA TAGATGTTAC ACCAACTTTG GAAGGGCAAA CATATTTCTT
5201 AATGAGAGGC AGTCCTTCAT GTTTTGCAAT AAAATGACTT TAAAAAAA
5251 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AA

```

25

BLAST Results

No BLAST result

30

Medline entries

35

No Medline entry

40

Peptide information for frame 3

ORF from 57 bp to 1187 bp; peptide length: 377

Category: putative protein

Classification: no clue

45

Prosites motifs: LEUCINE_ZIPPER (19-40)

50

```

1 MTDDSESVL SDSHEGSELE LPVIQLCGLV EELSYVNSAL KTETEMFEKY
51 YAKLEPRDQR PPRLSEIKIS AADYAQFRGR RRSKSRTGMD RGVGLTADQK
101 LELVQKEVAD MKDDL RHTRA NAERDLQHHE AIIEEAEIRW SEVSREVHEF
151 EKDILKAISK KKGSLATQK VMKYIEDMNR RRDNMKEKLR LKNVSLKVQR
201 KKMLLQLRQK EEVSEALHDV DFQQLKIENA QFLETIEARN QELTQLKLSS
251 GNTLQVLNAY KSKLHKAMEI YLNLDKIILL RKELLEKIEK ETLQVEEDRA
301 KAEAVNKRLR KQLAEFRAPQ VMTYVREKIL NADLEKSIRM WERKVEIAEM
55 351 SLKGHRKAWN RMKITNEQLQ ADYLAGK

```


.....

COILS

15

20 PS00029 37->59 LEUCINE_ZIPPER PDOC00029

25 HMM_NAME Helix-loop-helix DNA-binding domain

```

35  HMM                      AIEYIrsLQ*
                                IE  ++L+
Query                244 KIENAQFLE                252

```

DKFZphtes3_19p12

5 group: testis derived

DKFZphtes3_19p12 encodes a novel 664 amino acid protein without similarity to known proteins.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motif.

The new protein can find application in studying the expression profile of testis-specific genes.

15

unknown protein

Sequenced by MediGenomix

20

Locus: unknown

Insert length: 2161 bp

Poly A stretch at pos. 2086, no polyadenylation signal found

25

```
1  CCCGAGCCAG  CAACCCTGAG  GGGCGGCCGG  GCAGCGCCGC  CACCATGTTC
51  CTGGGCACCG  GGGAGCCGGC  CTTGGACACG  AGTCACCTTA  TCTCTCTAAG
101  CCGAGCGTCC  CTGACCCCTG  AGAAGCTGTG  GCTGGGAACC  GCAAAGCCAG
30  151  GAAGTCTGAC  CCAGGCCCTG  AACTCACCCC  TCACCTGGGA  GCATGCGTGG
201  ACTGGCGTCC  CCGGCGGCAC  TCCTGACTGT  CTGACAGACA  CTTTCAGAGT
251  GAAGAGGCCA  CATCTCAGGC  GCTCTGCCAG  CAACGGTCAT  GTCCCTGGGA
301  CTCCTGTCTA  CAGAGAAAAA  GAAGATATGT  ATGACGAGAT  TATTGAGTTA
35  351  AAGAAGTCAT  TGCACGTGCA  GAAGAGCGAC  GTGGACCTGA  TGAGAACGAA
401  GCTCCGGCGC  CTGGAGGAGG  AAAACAGCAG  GAAGGACCGG  CAGATAGAGC
45  451  AGCTCCTGGA  TCCCAGCCGC  GGCACGGATT  TTGTTCCGGAC  TCTGGCAGAG
501  AAAAGGCCCG  ATGCCAGTTG  GGTCAATTAAC  GGGCTGAAGC  AGAGGATCCT
551  GAAGCTGGAA  CAGCAGTGCA  AGGAGAAGGA  CGGCACCATC  AGCAAACCTCC
601  AGACCGATAT  GAAGACTACC  AACCTGGAAG  AGATGCGGAT  CGCCATGGAG
40  651  ACATACTACG  AGGAGGTGCA  TCGTCTCCAG  ACCCTCTTGG  CAAGTTCTGA
701  AACCACCGGA  AAGAAGCCCC  TGGGGGAGAA  GAAGACGGGC  GCCAAAAGGC
751  AGAAGAAGAT  GGGCAGTGCC  CTCCTGAGCT  TGTCCCGGAG  TGTCCAGGAG
801  CTCACGGAAG  AGAACCAGAG  CCTGAAGGAG  GACCTGGACC  GCGTGCTGAG
851  CACCTCCCCA  ACCATCTCCA  AGACACAGGG  TTATGTGGAG  TGGAGCAAGC
45  901  CCCGGCTGCT  GAGGCGCATT  GTGGAGCTGG  AGAAGAAACT  AAGTGTGATG
951  GAGAGCTCAA  AATCACACGC  CGCAGAGCCA  GTCAGATCAC  ACCCGCCAGC
1001  CTGCCTTGCA  TCCAGCTCTG  CGCTGCACAG  ACAGCCACGA  GGGGACCGCA
1051  ACAAGGACCA  CGAGCGTCTC  CGAGGGGCTG  TGAGAGACCT  GAAGGAAGAG
1101  CGGACCGCGC  TGCAGGAGCA  GCTGCTGCAG  AGAGATTTGG  AGGTGAAGCA
50  1151  GCTCCTGCAG  GCGAAGGCCG  ACCTGGAGAA  GGAGCTGGAG  TGC GCGAGGG
1201  AGGGCGAGGA  GGAGAGGAGA  GAGCGAGAGG  AGGTTTTGAG  AGAGGAGATT
1251  CAGACACTTA  CCAAGCAAGCT  CCAAGAATTG  CAAGAAATGA  AGAAAGAAGA
1301  GAAAGAGGAT  TGCCCCGGAAG  TTCCTCATAA  GGCCCCAAGAG  CTCCCAGCTC
1351  CCACTCCCA  CAGCAGGCAC  TGCAGCAAG  ACTGGCCGCC  GGATTCCAGC
55  1401  GAGGAGGGGG  TCCCGCGGCC  CCGCTCCCC  TGCTCTGATG  GGAGAAGAGA
1451  CGCCGCGGCC  AGAGTCCTGC  AGGCCAGTG  GAAGGTGTAC  AAGCACAAGA
1501  AAAAAAAGGC  TGTCTGGAT  GAGGCGGCTG  TGGTGCTTCA  GGCAGCTTTC
1551  AGGGGACATC  TCACGCGGAC  AAAGCTCTTA  GCAAGCAAAG  CACATGGCTC
```

1601 AGAGCCACCC AGCGTGCCAG GCCTCCCAGA CCAGAGCTCT CCTGTGCCCC
1651 GCGTTCGAG CCCATCGCC CAGGCCACGG GCAGCCCTGT GCAGGAGGAG
1701 GCCATCGTCA TCATCCAGTC CGCTCTGCGG GCACACCTGG CCCGGGCCAG
1751 GCACAGTGCT ACCGGTAAAA GAACCACCAC CGCAGCTTCT ACCAGGAGGA
5 1801 GATCGGCTTC AGCCACACAC GGGGACGCCT CCTCCCACC CTTCTCGCA
1851 GCTCTTCTTG ACCCTCTCTC CTCAGGGCCA CAGGCCTTGG CACCTCTACC
1901 TGGGGATGAC GTCAACTCCG ATGATTCCGA CGATATTGTC ATTGCACCGT
1951 CTCTGCCCAC GAAGAACTTT CCAGTTTAGG TCCCCGTCAC TGTCTCCACG
2001 CCGTGATGGC AGCGCTGCCG AGGACATAGG AACCACGACT GGAAAGATAA
10 2051 TTTATCGTGT TAGGAGAAGA ACGATGATAC CTACTTAAAA AAAAAAAAAA
2101 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
2151 AAAAAAAAAA A

15

BLAST Results

No BLAST result

20

Medline entries

No Medline entry

25

Peptide information for frame 3

30

ORF from 45 bp to 1976 bp; peptide length: 644
Category: similarity to unknown protein
Classification: unclassified
Prosites motifs: RGD (332-334)

35

1 MFLGTGEPAL DTSHLISLSR ASLTPQKLWL GTAKPGSLTQ ALNSPLTWEH
51 AWTGVPGGTP DCLTDTRVK RPHLRRSASN GHVPGETPVYR EKEDMYDEII
101 ELKKSLSHVQK SDVDLMRTKL RRLEEENSRL DRQIEQLLDP SRGTDVVRTL
40 151 AEKRPDASWV INGLKQRIK LEQQCKEKDG TISKLTQDMK TTNLEEMRIA
201 METYYEEVHR LQTLASSET TGKKPLGEKK TGAQRQKKMG SALLSLRSRV
251 QELTEENQSL KEDLDRVLST SPTISKTRQY VEWSPRLLR RIVELEKKLS
301 VMESKSHAA EPVRSHPPAC LASSSALHRQ PRGDRNKDHE RLRGAVRDLK
351 EERTALQEQ LQRDLEVQKL LQAKADLEKE LECAREGEEE RREREVLRE
45 401 EIQTLSKLQ ELQEMKKEEK EDCPEVPHKA QELPAPTSS RHCEQDWPPD
451 SSEEGLPRPR SPCSDGRRDA AARVLQAQWK VYKHKKKKAV LDEAAVVLQA
501 AFRGHLTRTK LLASKAHGSE PPSVPLPDQ SSPVPRVPSP IAQATGSPVQ
551 EEAIIVIIQSA LRAHLARARH SATGKRITTA ASTRRRSASA THGDASSPPF
601 LAALPDPSPS GPQALAPLPG DDVNSDDSDD IVIAPSLPTK NFPV

50

BLASTP hits

55 No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_19p12, frame 3

No Alert BLASTP hits found

Pedant information for DKFZphtes3_19p12, frame 3

5

Report for DKFZphtes3_19p12.3

10 [LENGTH] 644
 [MW] 71810.41
 [pI] 8.80
 [HOMOL] TREMBL:AB028946_1 gene: "KIAA1023"; product: "KIAA1023 protein"; Homo sapiens mRNA for KIAA1023 protein, partial cds. 0.0
 15 [FUNCAT] 30.03 organization of cytoplasm [S. cerevisiae, YDL058w] 2e-07
 [FUNCAT] 08.07 vesicular transport (golgi network, etc.) [S. cerevisiae, YDL058w] 2e-07
 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YLR309c] 3e-06
 20 [FUNCAT] 30.04 organization of cytoskeleton [S. cerevisiae, YDR356w] 2e-05
 [FUNCAT] 09.10 nuclear biogenesis [S. cerevisiae, YDR356w] 2e-05
 25 [FUNCAT] 03.22 cell cycle control and mitosis [S. cerevisiae, YDR356w] 2e-05
 [FUNCAT] 98 classification not yet clear-cut [S. cerevisiae, YJR134c] 4e-05
 [BLOCKS] DM013541
 30 [BLOCKS] BL006278 GHMP kinases ATP-binding domain proteins
 [BLOCKS] BL00326C Tropomyosins proteins
 [BLOCKS] BL011608 Kinesin light chain repeat proteins
 [BLOCKS] BL00820D Glucoamylase proteins region proteins
 [BLOCKS] BP04417C
 35 [BLOCKS] BL004128 Neuromodulin (GAP-43) proteins
 [EC] 3.6.1.32 Myosin ATPase 3e-08
 [PIRKW] tandem repeat 3e-08
 [PIRKW] transmembrane protein 2e-07
 [PIRKW] muscle contraction 3e-08
 40 [PIRKW] actin binding 3e-08
 [PIRKW] ATP 3e-08
 [PIRKW] thick filament 3e-08
 [PIRKW] alternative splicing 7e-07
 [PIRKW] coiled coil 3e-08
 45 [PIRKW] P-loop 3e-08
 [PIRKW] heptad repeat 2e-07
 [PIRKW] methylated amino acid 3e-08
 [PIRKW] hydrolase 3e-08
 [PIRKW] Golgi apparatus 2e-07
 50 [SUPFAM] myosin heavy chain 3e-08
 [SUPFAM] myosin motor domain homology 3e-08
 [SUPFAM] alpha-actinin actin-binding domain homology 8e-06
 [SUPFAM] plectin 8e-06
 [SUPFAM] ribosomal protein S10 homology 8e-06
 55 [SUPFAM] giantin 2e-07
 [PROSITE] RGD 1
 [KW] All_Alpha
 [KW] LOW_COMPLEXITY 14.60 %

[KW] COILED_COIL 15.22 %

5 SEQ MFLGTGEPALDTSHLISLSRASLTPQKLWLGTAKPGSLTQALNSPLTWEHAWTGVPGGTP
SEG
PRD ccc
COILS
10 SEQ DCLTDTFRVKRPHLRRSASNGHVPGTPVYREKEDMYDEIIELKKS LHVQKSDVDLMRTKL
SEG
PRD cccccchhh
COILScc
15 SEQ RRLEEENS RKDRQIEQLLDPSRGTD FVRTLA EKRPDASWVINGLKQRILKLEQ QCKEKDG
SEG
PRD hhh
COILS
20 CCCCCC.....
SEQ TISK LQTD MKTTNLEEMRIAMETYEEVHRLQTLLASSETTGKKPLGEKKTGAKRQKKMG
SEG
PRD hhh
COILSCCCC
25
SEQ SALLSLSRSVQELTEENQSLKEDLDRVLSTSP TISK TQGYVEWSKPRLLRRIVELEKKLS
SEG
PRD hhh
COILScc
30
SEQ VMES SKSHAAEPVR SHPPAC LASSSALHRQPRGDRNKDHERLRGAVRDLKEERTALQEQ L
SEG
PRD hhh
COILS
35
40 SEQ LQRDLEV KQLLQAKADLEKELE CAREGEEERRERE EVLREEIQTL TSKLQELQEMKKEEK
SEGxx
PRD hhh
COILScc
45
SEQ EDCPEVPHKAQELPAPTSSRHCEQDWPPDSSE EGLPRPRSPCSDGRRDAAARVLQ AQWK
SEG x.....x
PRD hhh
COILS
50 CCCCCC.....
SEQ VYKHKKKKAVLDEAAVVLQAAFRGHLTRTKLLASKAHGSEPPSVPGLPDQSSPVPRVPSP
SEG xxxxxxxx.....
PRD hhh
COILS
55
SEQ IAQATGSPVQEEAIVIIQ SALRAHLARARHSATGKRTTTAASTRRRRSASATHGDASSPPF

SEGxx.....
 PRD cccccccccceeehhhhhhhhhhhhhhhhhhccccceeehhhhhhhhhhccccccccccce
 COILS

5

SEQ LAALPDPSPSGPQALAPLPGDDVNSDDSDDIVIAPSLPTKNFPV
 SEGxxxxxxxxxxxxxxxx.....
 PRD eeeeeccccccccccccccccccccccccccccceeeeccccccccccc
 COILS

10

Prosites for DKFZphtes3_19p12.3

15

PS00016 332->335 RGD PD0C00016

(No Pfam data available for DKFZphtes3_19p12.3)

- 5 group: transmembrane protein
- DKFZphtes3_20h12 encodes a novel 1204 amino acid protein without similarity to known proteins.
- 10 The novel protein contains 1 transmembrane region and two leucine zippers.
No informative BLAST results; No predictive prosite, pfam or SCOP motife.
- 15 The new protein can find application in studying the expression profile of testis-specific genes and as a new marker for testicular cells.
- 20 putative protein
- perhaps complete cds.
Pedant: TRANSMEMBRANE 1
- 25 Sequenced by MediGenomix
- Locus: unknown
- Insert length: 5894 bp
- 30 Poly A stretch at pos. 5874, no polyadenylation signal found

```

      1 CTCTGCCTTT CCTCTCGCAG CCACCCCTTCC TCTCAGACCA GTACGGTGGC
    51 CGACGGGAGT CAGACGCTGG GGATGAATGA AGGATCAACA AACAGTAATA
  35 101 ATGACTGAAT GTACAAGTCT TCAGTTTGTC AGCCCTTTTG CTTTTGAGGC
    151 AATGCAGAAG GTGGATGTTG TTTGCCTGGC ATCTTTAAGT GATCCAGAAT
    201 TAAGACTTCT TCTGCCCTGT TTGGTACGGA TGGCACTTTG TGCACCTGCT
    251 GACCAGAGCC AAAGCTGGGC TCAGGATAAG AAACATCATCC TTCGCCTTCT
    301 TTCTGGAGTG GAAGCTGTCA ACTCCATTGT TGCATTGTTG TCCGTGGACT
  40 351 TTCATGCTTT AGAACAAGAT GCCAGCAAAG AACAGCAGCT TAGGCATAAA
    401 CTTGGAGGAG GCAGTGGAGA GAGCATCCTG GTATCACAGC TTCAGCATGG
    451 ACTGACGTTA GAGTTTGAAC ACAGTGATTC ACCTCGTCGA TTGCGTCTTG
    501 TGCTTAGTGA ACTGTTGGCA ATTATGAACA AGGTGTCTGA GTCCAACGGA
    551 GAATTTTTTT TCAAGTCTTC TGAACTTTTT GAGAGTCCAG TATATTTGGA
  45 601 GGAAGCTGCA GATGTACTTT GTATTTTACA AGCAGAGCTC CTTTCCTTGC
    651 TCCCTATAGT TGATGTAGCT GAAGCTTTGC TACATGTTAG AAATGGTGCC
    701 TGGTTCTTGT GTCTCTTGGT GGCCAATGTT CCTGATAGTT TTAATGAAGT
    751 TTGTAGGGGC CTGATAAAAA ATGGAGAACG ACAAGATGAA GAAAGTCTTG
    801 GAGGAAGGCG CAGGACAGAT GCCTTACGCT TCTTGTGTAA AATGAATCCT
  50 851 TCTCAGGCCC TCAAGGTCCG AGGCATGGTG GTGGAAGAAT GTCACTTGCC
    901 AGGCCTTGGT GTGGCTTTGA CATTGGATCA TACTAAAAAT GAAGCTTGTG
    951 AGGATGGAGT GAGTGACTTG GTTTGTTTTG TAAGTGGTTT GCTTCTTGGA
  1001 ACAAATGCGA AAGTCCGGAC TTGGTTTGGG ACTTTTATCC GAAATGGACA
  1051 GCAGAGAAAA AGAGAGACCA GCAGTCTGT CTTTGGCAG ATGAGAAGGC
  55 1101 AGCTTCTTCT GGAGTTGATG GGCATTCTTC CCACAGTAAG AAGCACCGA
  1151 ATTGTGGAAG AAGCTGATGT GGATATGGAG CCCAATGTGT CTGTGTATTG
  1201 GGGGCTGAAA GAAGAGCATG TTGTGAAAGC CAGTGCCTC TTACGTCTGT
  1251 ACTGTGCTTT GATGGGGATC GCTGGACTCA AACCAACTGA AGAAGAAGCT

```

1301 GAGCAATTAC TGCAGTTGAT GACGAGCCGT CCTCCTGCTA CGCCAGCTGG
1351 GGTTTCGCTT GTTTCACTTT CCTTTTGTAT GCTACTGGCC TTTTCTACAC
1401 TTGTCAGTAC ACCTGAACAG GAGCAGCTGA TGGTGGTGTG GCTAAGTTGG
1451 ATGATAAAAG AAGAAGCGTA TTTTGAGAGT ACTTCAGGCG TCTCTGCTTC
5 1501 TTTTGGGGAG ATGTTATTAT TGGTGGCTAT GTACTTTCAC AGCAACCAGC
1551 TTAGTGCTAT CATTGACTTG GTCTGTTCCA CTTTGGGGAT GAAGATTGTA
1601 ATTAAGCCAA GCTCCTTGAG CAGGATGAAG ACAATCTTCA CACAGGAAAT
1651 TTTTACTGAG CAGGTTGTCA CAGCTCATGC AGTTCGGGTC CCTGTCACCA
1701 GCAACCTGAG TGCCAACATT ACTGGATTTT TGCCTATTCA TTGTATTTAC
10 1751 CAGCTTCTCA GGAGCCGTTT CTTTACCAAG CACAAAGTGT CAATAAAAGA
1801 TTGGATTTAT AGACAGCTGT GTGAAACCTC TACTCCACTT CATCCTCAAT
1851 TACTTCCCTT GATTGATGTG TACATAAATT CTATACTTAC TCCTGCGTCTG
1901 AAATCTAATC CAGAAGCCAC AAATCAGCCA GTCACAGAAC AGGAGATACT
1951 CAATATTTTC CAAGGAGTCA TTGGGGGTGA CAACATCCGC CTTAATCAGC
15 2001 GTTTCAGTAT CACAGCACAG CTTTTGGTGC TCTACTATAT ACTGTCTTAT
2051 GAAGAGGCTC TTCTAGCAAA CACGAAGACT TTAGCTGCCA TGCAAAGAAA
2101 GCCCCAATCA TATTCTTCTT CTTTAATGGA TCAGATTCTT ATCAAATTCC
2151 TTATTCGACA GGCTCAAGGG CTGCAGCAGG AGTTGGGAGG GTTGCATTCA
2201 GCTTTACTAC GTCTCCTTGC TACTAACTAC CCACATTTAT GTATTGTGGA
20 2251 TGACTGGATT TGTGAAGAAG AAATCACAGG GACTGATGCC CTGCTACGGC
2301 GAATGCTCCT GACTAATAAT GCTAAAAATC ATTCTCCCAA ACAACTCCAA
2351 GAAGCATTTT CAGCTGTCCC AGTAAATCAC ACACAAGTGA TGCAGATTAT
2401 AGAACACTTG ACTCTACTCT CTGCCAGTGA ACTTATACCA TATGCGGAAG
2451 TGTTAACATC CAATATGAGC CAGCTATTGA ATTCAGGGGT TCCACGGAGA
25 2501 ATTCTGCAAA CAGTCAATAA ACTATGGATG GTTCTTAATA CTGTGATGCC
2551 TAGAAGGCTA TGGGTAATGA CGGTTAATGC ACTTCAGCCT TCAATAAAGT
2601 TTGTACGACA ACAAAGTAT ACTCAGAATG ACCTGATGAT AGATCCTCTC
2651 ATTGTCTTAA GGTGTGATCA GAGGGTTTCA AGATGCCCCC CACTGATGGA
2701 TATTACCCTA CACATGTTGA ATGGATATCT TCTTGCTATCT AAAGCCTACC
30 2751 TTAGTGCTCA TCTGAAGGAA ACAGAGCAAG ATAGGCCTTC CAGAATAAT
2801 ACAATTGGTT TAGTTGGACA AACTGATGCT CCGGAAGTTA CCAGGGAAGA
2851 ATTGAAAAAT GCATTACTGG CCGCTCAGGA TAGTGCAGCT GTCCAGATTCT
2901 TCTTAGAGAT TTGCCCTACCT ACTGAAGAGG AGAAAGCAAA TGGTGTCAAT
2951 CCAGATAGCT TGTTAAGAAA TGTTCAAAGT GTTATTACCA CCAGCGCTCC
35 3001 AAATAAGGGA ATGGAGGAAG GAGAAGACAA TTTGCTCTGT AACCTTCGAG
3051 AAGTTCAGTG CTTTATCTGT TGTCTCTTGC ACCAAATGTA CATTGCAGAT
3101 CCCAACATTG CTAAGCTTGT TCACTTTTCA GGTATCCAT GTGAAGTTTT
3151 GCCTCTGACG GTCGCAGGTA TTCCATCTAT GCACATCTGT CTAGATTCTCA
3201 TACCTGAGCT TATTGCACAG CCAGAACTTG AGAAACAGAT ATTTGCTATC
40 3251 CAGTTGCTTT CTCACTTGTG TATACAATAT GCATTACCAA AGTCACTTAG
3301 TGTGGCTCGT TTAGCTGTCA ATGTCATGGG AACTTTGTTA ACAGTTTTAA
3351 CACAGGCTAA GCGGTATGCT TTTTTTATGC CAACTCTGCC AAGTTTGGTC
3401 TCTTTTTGTC GAGCATTTC TCCATTGTAT GAGGATATTA TGTCTTTGCT
3451 GATCCAAATA GGGCAAGTTT GTGCCCTCTGA TGTGCGCACT CAGACAAGAG
45 3501 ACATTGATCC AATTATTACA CGTCTTCAAC AAATAAAGGA GAAACCAAGT
3551 GGATGGTCTC AAATCTGTAA AGATTCTCT TATAAAAATG GATCCAGGGA
3601 CACTGGAAGC ATGGATCCTG ATGTACAGCT CTGTCACTGT ATTGAAAGAA
3651 CAGTAATTGA AATAATAAAT ATGAGTGTTA GTGGAATTTA AAACAAAATT
3701 TAAAAACAACA AAAAGTTGTT TGCTGCATAT ACCCAACATG AATCTGCATA
50 3751 TTAGTAACAA CTCTAAACTG AATGGGAACA GTAAAGTATT GTCTTGGAAT
3801 CACTAAAACA ATTCAATTCA ACATGAGTAT AGTTTAGAAC TTTATGAGAA
3851 TTATGCTTGC TTGTTTCTGA TTGGCACATC TTTGGATCTA CTTTGCTGAT
3901 ATGTTTCTAT TGTAGCAGCT GAGCTTTTTT TTTTCCACT GGGAACACAT
3951 GTAAGAAACT CATTATTGGA AAGGGAATTT GGCCTTGTAT TTAGCTTTTG
55 4001 AAGTGAAGAC TGCCATGCCT TTAATTTCTT ATAAAAATGA GTCTGTGGGT
4051 AGCCCTAGTG TTTATTTTAA CTGTGAGCTT GTAACAGAAT GTGACAAAGA
4101 TGCAAAGATG GGAGAGGAAA AAAGGGTAAA GGGAAAGGAG AATTAAGGAA
4151 ATAATAGGAG TTA AAAACAC AAGTAGAAAT CTCAAAGATT TGCAGTGCAA

```

4201 GTAATAGTAA TGCAAGTTGG AATTCTAGTT CTCAAGAAAG AGTATTGAGA
4251 AGACTTTTAA AAAGGCAAGT AGCTTTTGTA AATGATTTCT GTGGAAATAC
4301 AGATGAGGAT TTAAAGATTT CACATATTTG CTTCAATTTT TATTAATATA
4351 TGAAGCCATA TGTTTAAAGA GATACTTGAA TAATTTGGAA TTTTAAGATA
5 4401 CTGGTGTAAG AGTGTTTACA GAAACATCTT TGTTCAAAGA AGAACCTGAG
4451 AGATCTCATT TAGTTTTATG TTTTAAATTT ATTTTATAA TGCTTTATTA
4501 ACTTACCTAA TGCTCAGAGG GGGGAAATAT GTATCAAATT AAATGAAGGT
4551 AGAGCAATAA AACCCTACTG ATTAAGAGAG TCTTGGTTTG TCATCAGGAT
4601 TATAATTCAT ATCTTACTTT GAGAAGATCT TTGAGTAAGA AAATGCAGTG
10 4651 TTTGAACCTG AGGAAAAGTT AAAGTGTAGA AAATATTGTC TTGCCGAAGG
4701 ATTTTGCAAG CCTCTGTGAG TAACTTCCAT TGATTAGGCA GACATATTCA
4751 GGTAACCCCT AATCATTAAA AAAAAATTAT CAATGTAGAA AGTAATTCCT
4801 TTTTTTCTCT CTGAGATATA CCTCAATCAC ACACCTCCCC ACCCCCACTT
4851 GAAACAGACC TCTTCACTTG TGTTTTTTTT TTTTTTTTCC TGAGGTGGAG
15 4901 TCTTCCCCTG TTGCCCAGGC TGGAGTGCAG TGGGATGATC TTGGCTCACT
4951 GCAACTTCTG CCACCTGGGT TCAAGGGATT CTCGTGCCTC AACCTCCTGA
5001 GTAGCTGGGA CTGCAGGCAC GCGCCACCTG TATTTTTGTA TTTTGTAGTAG
5051 AGACGGGGGT TTGCCATGTT GCCCAGACTG GTTTTGAACCT CCTGGCCTCA
5101 GGTGATCTGC CCACCTTGGC CTCCCAAAGT GCTGGGATTA CAGGTGTGAG
20 5151 CCACCGCACC TGGCCAGACC GCTTCACTTG TAAAAGAAAT TAGGCTAATA
5201 AGAAGGTGTA GTTTTTGAGA AATGAAATTT AACTTTAGCC TTTTCACTAG
5251 TAAATAGTCA CATCTCATTT TCTTCTTTTG TAAAATGGGG TTTACTACTGG
5301 CCCTACCTCA TATTCTATGA GAATGAGTTT GTAGCTGTTT CAAATCATGA
5351 AGTGATAGT ATCACATGTG ATAGAATATT TATAACTTTT TATTAGATGC
25 5401 TTAATGTTCA ATTAAGTAAT TTTGATGTGA AAAATAAAAG TAATAAAAGT
5451 ATCTTAAAAA TAGCATAAGA ATTTTCATAT TTTTAAACAA GGCAGTTTTG
5501 TAGTCCCTTA AGATTAAATA CAACTGCTCC TTTTTTTTTT AAAGTGGGC
5551 CTTGCGATAT TTTGTGTGAA TAGATATGCC CTAGGAGTTC AGAAAAAGTT
5601 AAAAGTATGT TTTCTAATTA AATGCAGTGC ACATTCCTGG ATCAATATTC
30 5651 AAAGACTGGT CATAACCTGC TGTGTTAAAA TAATCACATA TGCTCTTTTT
5701 CATCAGATTT GTTGATGATG TAAATAAAAT GTGTAAATAT ATTAGTAAAT
5751 GTTAATATTC ATGTATTTTA AGTTAAGGTT ATAAAATTTG TCACAATGTG
5801 TTTTTTTATT CAAGTGAAAA CAGATGTGTG CAGCTATTTT GAATATTGGT
5851 TTATAAACAT TCATATTCTT TATCAAACAA AAAAAAAAAA AAAA

```

BLAST Results

40 No BLAST result

Medline entries

45 No Medline entry

50 Peptide information for frame 2

ORF from 77 bp to 3688 bp; peptide length: 1204

Category: putative protein

55 Classification: unclassified

Prosites motifs: LEUCINE_ZIPPER (167-184)

LEUCINE_ZIPPER (692-709)

```

1 MKDQQTVIMT ECTSLQFVSP FAFEAMQKVD VVCLASLSDP ELRLLLPCLV
5 51 RMALCAPADQ SQSWAQDKKL ILRLLSGVEA VNSIVALLSV DFHALEQDAS
101 KEQQLRHKLQ GSGESILVS QLQHGTLTLEF EHS DSPRRLR LVLSELLAIM
15 151 NKVSENGEF FFKSSELFES PVYLEEAADV LCILQAE LPS LLPIVDVAEA
201 LLHVRNGAWF LCLLVANVPD SFNEYCRGLI KNGERQDEES LGGRRRTDAL
251 RFLCKMNPSQ ALKVRGMVVE ECHLPGLGVA LTL DHTKNEA CEDGVSDLVC
301 FVSGLLLG TN AKVRTWFGTF IRNGQQRKRE TSSSVLWQMR RQLLLELMGI
351 LPTVRSTRIV EEADVDMEPN VSVYSGLKEE HVVKASALLR LYCALMG IAG
10 401 LKPTEEEEAEQ LLQLMTSRPP ATPAGVRFVS LSFCMLLAFS TLVSTPEQEQ
451 LMVVWLSWMI KEEAYFESTS GVSASFGEML LLVAMYFHSN QLSAII DLVC
501 STLGMKIVIK PSSLSRMKTI FTQEIFTEQV VTAHAVRVPV TSNLSANITG
551 FLPIHICIYQL LRSRSFTKHK VSIKDWIYRQ LCETSTPLHP QLLPLIDVYI
601 NSILTPASKS NPEATNQPV T EQEILNIFQG VIGGDNIRLN QRF SITAQLL
15 651 VLYYILSYEE ALLANTKT LA AMQRKPKSYS SSLMDQIPIK FLIRQAQGLQ
701 QELGGLHSAL LRLLATNYPH LCIVDDWICE EEITGTDALL RRMLLTNNAK
751 NHSPKQLQEA FSAVPVNHTQ VMQIIIEH TL LSASELIPYA EVLTSNMSQL
801 LNSGVPRRIL QTVNKLWMLV NTVMPRRLWV MTVNALQPSI KFVRQ QKYTQ
851 NDL MIDPLIV LRCDQRVHRC PPLMDITLHM LNGYLLASKA YLSAHLKETE
20 901 QDRPSQNN TI GLVGQTDAP E VTREELKNAL LAAQDSA AVQ ILLEICLPTE
951 EEKANGVNP D SLLRNVSQVI TTSAPNKGME EGEDNLLCNL REVQCLICCL
1001 LHQMYIADPN IAKLVHFQGY PCELLPLTVA GIPSMHICLD FIPELIAQPE
1051 LEKQIFAIQL LSHLCIQYAL PKSLSVARLA VNVMG TLLTV LTQAKRYAFF
1101 MPTLPSLV SF CRAFPPLYED IMSLLIQIGQ VCASDVATQT RDIDPIITRL
25 1151 QQIKEKPSGW SQICKDSSYK NGSRD TGSM D PDVQLCHCIE RTVIEIINMS
1201 VSGI

```

30 BLASTP hits

No BLASTP hits available

35 Alert BLASTP hits for DKFZphtes3_20h12, frame 2

No Alert BLASTP hits found

Pedant information for DKFZphtes3_20h12, frame 2

40 Report for DKFZphtes3_20h12.2

```

45 [LENGTH] 1204
[MW] 134347.53
[pI] 5.75
[MOLE] TREMBL:CEZC37b_3 gene: "ZC37b.b"; Caenorhabditis
elegans cosmid ZC37b 2e-22
[PROSITE] LEUCINE_ZIPPER 2
50 [KW] TRANSMEMBRANE 1
[KW] LOW_COMPLEXITY 2.57 %
[KW] COILED_COIL 2.33 %

55 SEQ MKDQQTVIMTECTSLQFVSPFAFEAMQKVDVVCLASLSDPELRLLLPCLV
SEG .....
PRD cccccccccccccccccchhhhhhhhhheeeeeccccchhhhhhhchhhhhhhccccc

```

COILS

MEM

5 SEQ SQSWAQDKKLILRLLSGVEAVNSIVALLSVDFHALEQDASKEQQLRHKLGGGSGESILVS
SEG
PRD hhhhhhhhhhhhhhhhhccccccccccccccccchhhhhhhhhhhhhhhhhccccceeeec
COILS

10 MEM

SEQ QLQHGTLLEFEHSDSPRRLRLVLSELLAIMNKVSESNGEFFFKSSELFESPVYLEEAADV
SEGxxxxxxxxxxxxx.....
PRD cccccceeeecccccchhhhhhhhhhhhhhhhhhhccccccccccccccccchhhhhhhhh
COILS

15 MEM

20 SEQ LCILQAEPLSLLPIVDVAEALLHVRNGAWFLCLLVANVPDSFNEVCRGLIKNGERQDEES
SEG
PRD hhhhhhccccchhhhhhhhhhhhhhhccchhhhhheeeccccccchhhhhcccccccccccc
COILS

25 MEM

SEQ LGGRRRTDALRFLCKMNPSQALKVRGMVVEECHLPGLGVALTLDHTKNEACEDGVSDLVC
SEG
PRD ccccchhhhhhhhhhhccccceeeeeeeeeeeeeccccccccceeeccccccccccccccccceee
COILS

30 MEM

35 SEQ FVSGLLLGTNAKVRTWFGTFIRNGQQRKRETSSSVLWQMRRQLLLELMGILPTVRSTRIV
SEG
PRD eeccccccccceeeeeeeeeeeeeecchhhhhccccchhhhhhhhhhhhhhhhhccccceeeeee
COILS

MEM

40 SEQ EEADVDMEPNVS SVYSGLKEEHVVKASALLRLYCALMGIAGLKPTEEAEQLLQLMTSRPP
SEG
PRD eeccccccccceeeccccchhhhhhhhhhhhhhhhhhhhhccchhhhhhhhhhhhhhhcccc
COILS

45 MEM

SEQ ATPAGVRFVSLSF CMLLAFSTLVSTPEQEQLMVVWLSWMIKEEAYFESTSGVSASFGEML
SEG
PRD cccccceeeehhhhhhhhhhhccccccchhhhhhhhhhhhhhhhhhhhhccccccccchhhhhh
COILS

50 MEMMMMMMMMMMMMMMMMM.....

55 SEQ LLVAMYFHSNQLSAIIDLVCSTLGMKIVIKPSSLSRMKTIFTQEIFTEQVVTAAHAVRVPV
SEG
PRD hhhhhhhccchhhhhhhhhhhhhccccceeeecccccchhhhhhhhhhhhhhhhhhhheec
COILS

.....

MEM
SEQ TSNLSANITGFLPIHCIYQLLRSRSFTKHKVSIKDWIYRQLCETSTPLHPQLPLIDVYI
SEG
5 PRD cccccceeeeeehhhhhhhhhhhhhhhccccccchhhhhhhhhccccccccccccceeee
COILS
MEM
10 SEQ NSILTPASKSNPEATNQPVTEQEI LNIFQGVIGGDNI RLNQ RFSITAQLLVLYI LS YEE
SEG
PRD eccccccccccccccccchhhhhhhhhhhhhhhccccccceeeehhhhhhhhhhhhhhhhhhhhh
COILS
15 MEM
SEQ ALLANTKT LAAMQ RKPKSYSSSLMDQIPIKFLIRQAQGLQELGGLHSALLRLLATNYPH
SEGxxxxxxxxxxxxxxxxxxxxx.....
20 PRD hhhhhhhhhhhhhhhccccccccccccchhhhhhhhhhhhhhhhhccccchhhhhhhhhcccc
COILSCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC.....
MEM
25 SEQ LCIVDDWICEEEITGTDALLRRMLLTNNAKNHSPKQLQEA FSAVPVNHTQVMQIEHLTL
SEG
PRD eeeeeceeeeechhhhhhhhhhhhhhhccccccccchhhhhhhhhhhccccchhhhhhhhhhhhh
COILS
MEM
30 SEQ LSASELIPYAEVLTSNMSQLLNSGVPRRILQTVNKLWMVLNTVMPRRLWVMTVNALQPSI
SEG
PRD hhhhhhhhhhhccccccchhhhhccccchhhhhhhhhhhhhhhhhccccchhhhhhhccccch
COILS
35 MEM
SEQ KFVRQKQYTQNDLMIDPLIVLRCDQ RVHRC PPLMDITLHMLNGYLLASKAYLSAHLKETE
SEG
40 PRD hhhhhhhccccccccccccceeececcccccccccccceeeccccccchhhhhhhhhhhhhhhhh
COILS
MEM
45 SEQ QDRPSQNN TIGLVGQTD APEVTREELKNALLAAQDSA AVQILLEICLPTEEEKANGVNP
SEG
PRD cccccccccccccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhcccccccccccccc
COILS
50 MEM
SEQ SLLRNVQSVITTSAPNKGMEEGEDNLLCNLREVQCLICLLHQMYIADPNIAKL VHFQGY
SEG
55 PRD cceeeeeeeeececcccccccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhccccceeecccc
COILS
MEM

PCT/IB01/02050

```
5  MEM .....
```

```

10 PRD  hhhhhhhhhhhhhhhhhhhhhccccceeeccccccchhhhhhhhhhhhhcchhhhhcccc
    COILS
    .....
    MEM .....

```

20 MEM

25 COILS
MEM

PS00029	167->189	LEUCINE_ZIPPER	PD0C00029
PS00029	692->714	LEUCINE_ZIPPER	PD0C00029

-344-

DKFZphtes3_21k14

5 group: testis derived

DKFZphtes3_21k14 encodes a novel 558 amino acid protein without similarity to known proteins.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motif.

The new protein can find application in studying the expression profile of testis-specific genes.

15

unknown protein

perhaps complete cds.

20

Sequenced by LMU

Locus: unknown

25 Insert length: 2547 bp

Poly A stretch at pos. 2506, polyadenylation signal at pos. 2479

```

    1  GGCCACGTTT  AGCGGACACG  GGAGCAAGAT  GGCGATTCCG  GGCAGGCAGT
30    51  ATGGGCTTAT  TTTGCCAAAG  AAAACACAGC  AGTTGCACCC  TGTTTTGCAA
   101  AAACCATCAG  TGTTTGGGAA  TGATTCTGAT  GATGATGATG  AGACCTCTGT
   151  GAGTGAAAGC  CTTCAGAGGG  AAGCTGCTAA  GAAGCAGGCC  ATGAAACAGA
   201  CCAAACCTGA  AATCCAGAAG  GCCCTTGCA  AAGATGCTAC  TGTGTATGAA
   251  TATGACAGTA  TTTATGATGA  AATGCAGAAA  AAAAAGGAGG  AAAATAATCC
35   301  CAAATTGCTT  TTGGGGAAAG  ACAGAAAGCC  CAAGTATATT  CACAACCTTG
   351  TAAAAGCAGT  TGAGATCAGA  AAAAAGGAAC  AGGAAAAAAG  AATGGAAAAA
   401  AAAATACAGA  GAGAACGAGA  AATGGAAAAA  GGGGAGTTTG  ATGATAAAGA
   451  AGCATTGTG  ACATCTGCAT  ATAAGAAAAA  ACTGCAAGAG  AGAGCTGAAG
   501  AAGAAGAAAG  AGAAAAGAGG  GCTGCTGCAC  TGGAAGCATG  TTTGGATGTA
40   551  ACCAAGCAGA  AAGATCTCAG  TGGATTTTAT  AGGCACCTAT  TAAATCAAGC
   601  AGTTGGTGAA  GAGGAAGTAC  CTAAATGCAG  CTTTCGTGAA  GCCAGATCTG
   651  GTATAAAGGA  AGAAAAATCA  AGGGGCTTCT  CCAATGAAGT  AAGTTCAAAA
   701  AACAGAATAC  CACAAGAGAA  ATGCATTCTT  CAAACTGATG  TGAAAGTAGA
   751  GGAAAACCCA  GATGCAGACA  GTGACTTCGA  TGCTAAGAGC  AGTGCGGATG
45   801  ATGAAATAGA  AGAACTAGA  GTGAACTGCA  GAAGGGAAAA  GGTCATAGAG
   851  ACCCTGAGA  ATGACTTCAA  GCACCACAGG  AGTCAAAACC  ACTCTCGGTC
   901  ACCTAGTGAA  GAAAGAGGGC  ACAGTACCAG  GCACCACACG  AAAGGATCAC
   951  GAACGTCGAG  AGGACATGAG  AAAAGGGAAG  ATCAGCACCA  GCAGAAGCAA
50  1001  TCCAGAGACC  AAGAGAACCA  TTACACTGAC  CGTGATTACC  GGAAAGAAAG
  1051  GGATTCTCAT  AGGCACAGAG  AGGCCAGTCA  TAGAGATTCC  CATTGGAAGA
  1101  GGCATGAACA  GGAAGATAAA  CCAAGGGCGA  GGGACCAAAG  AGAAAGAAGT
  1151  GACAGAGTAT  GGAAAAGGGA  GAAAGATAGG  GAGAAATATT  CCCAAAGAGA
  1201  ACAAGAAAGA  GATAGACAAC  AAAATGATCA  GAACCGACCC  AGTGAGAAAG
  1251  GAGAGAAGGA  AGAGAAAAGG  AAAGCAAAGG  AAGAGCATAT  GAAAGTAAGG
55  1301  AAGGAAAGAT  ATGAAAATAA  TGATAAATAC  AGAGATAGAG  AAAAACGAGA
  1351  GGTAGGTGTT  CAGTCTTCAG  AAAGAAATCA  AGACAGAAAG  GAAAGACGCC
  1401  CAAATTCTAG  GGCAAAGGAT  AAATTTCTTG  ACCAAGAAAG  ATCCAACAAA
  1451  ATGAGAAACA  TGGCAAAGGA  CAAAGAAAGA  AACCAAGAGA  AACCCTCTAA
```

```

1501 TTCTGAATCA TCACTGGGAG CAAAACACAG ACTCACAGAG GAAGGGCAAG
1551 AGAAGGGTAA AGAACAAGAG AGACCACCTG AGGCAGTGAG CAAGTTTGCA
1601 AAGCGGAACA ATGAAGAAAC TGTAATGTCA GCTAGAGACA GGTACTTGCC
1651 CAGGCAGATG GCGCGGGTTA ATGCAAAGAC CTATATTGAG AAAGAAGATG
5 1701 ATTGATGGCT ACCCCAAGAG AAAGATTTAA GGAAGCACAG AAAACTGTAA
1751 TTCCTGGAAC CTGCTGCGTA AAACCATAAA GGAGTGTGTT ACCAGTAGTT
1801 TGGAGGGCAT TTTTAAATTT ATTTTCAAAA TTTTAAGTTA AAAGTCAGTC
1851 TTACAGCTTG GATGTTTGGG TGTGGATGTT TGGCTGAATT TATATATAGT
1901 GTGTACTCAT CAATACCACA TTCTTTGTTG TATTCAAGAA CCGTTAAGAG
10 1951 TGTGCTAATT CCCTGTAGGT ACATAATGAG GAAAATTTGC TCCACTACAA
2001 CCATTAATAA ATAATTTTGG CCAGATACGG TAGCTCGTGC CTGTAATACC
2051 AACATTTTGG GAGGCCAAGG CAGAAGGATA TTGAGGCTAG GCATTCAAGA
2101 CCAGCCTAGG CAGGATAATA AGACCTTGTC TCTATTTAAA AAACAAAAAG
2151 CCTAGCATGG TAGTCCATGC CTGTAGTCCC AGCTGTTTGA GAGGCTGAGG
15 2201 CAAGAAGATC ACTTGAGCCT AGGAATTTGA TGTTACAGTG AGGTATGATC
2251 ATGCCACTGC ACTCCAACCT GGGCAACAGA ATGAGACCCT GTCTCTAAAA
2301 AATTTTTTTT AAATAAATAA TTTAACTCTT CTAATAATGT TTTGTTGCGAG
2351 GAAATGTATT TCAGATAAAA TATGGATTTG AAAAACAGAA AATATACTTT
2401 ATGTTCTGAA ATTTGTATTT AAGTATAAAA TGTGAATCAT CTTGTCTAAA
20 2451 TAGCTTACAG CATAGTTGGC TTAAATGAAA ATAAAATGAT ATGCTTATAC
2501 ATTTGGAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAG

```

BLAST Results

25

No BLAST result

30

Medline entries

No Medline entry

35

Peptide information for frame 2

```

40 ORF from 29 bp to 1702 bp; peptide length: 558
Category: similarity to unknown protein
Classification: Nucleic acid management

```

```

1 MAIPGRQYGL ILPKKTQQLH PVLQKPSVFG NDSDDDDETS VSESLQREAA
45 51 KKQAMKQTKL EIQKALAEDA TVYEYDSIYD EMQKKKEENN PKLLLKGDRK
101 PKYIHNLLKA VEIRKKEQEK RMEKKIQRRER EMEKGEFDDK EAFVTSAYKK
151 KLQERAEEEE REKRAAALEA CLDVTKQKDL SGFYRHLLNQ AVGEEVVKC
201 SFREARSGIK EEKSRGFSNE VSSKNRIPQE KCILQTDVKV EENPDADSDF
251 DAKSSADDEI EETRVNCRRE KVIETPENDF KHHRSQNHRS SPSEERGHST
50 301 RHHTKGSRTS RGHEKREDQH QQKQSRDQEN HYTD RDYRKE RDSHRHREAS
351 HRD SHWK RHE QEDKPRARDQ RERSDRVWKR EKDRKYSQR EQERDRQQND
401 QNRPSEKGEK EEKSKAKEEH MKVRKERYEN NDKYRDREKR EVGVQSSERN
451 QDRKESSPNS RAKDKFLDQE RSNKMRNMAK DKERNQEKPS NSESSLGAKH
501 RLTEEGQEKG KEQERPPEAV SKFAKRNNNEE TVMSARDRYL ARQMARVNAK
55 551 TYIEKEDD

```

BLASTP hits

No BLASTP hits available

5 Alert BLASTP hits for DKFZphtes3_21k14, frame 2

No Alert BLASTP hits found

10 Pedant information for DKFZphtes3_21k14, frame 2

Report for DKFZphtes3_21k14.2

15 [LENGTH] 567
 [MW] 67262.89
 [pI] 8.96
 [HOMOL] TREMBL:AC006233_14 gene: "F12K2.14"; Arabidopsis
 thaliana chromosome II BAC F12K2 genomic sequence, complete
 20 sequence. 3e-11
 [FUNCAT] 04.99 other transcription activities [S. cerevisiae,
 YKR092c] 1e-05
 [FUNCAT] 30.10 nuclear organization [S. cerevisiae, YKR092c]
 1e-05
 25 [FUNCAT] 06.07 protein modification (glycosylation, acylation,
 myristylation, palmitylation, farnesylation and processing)
 [S. cerevisiae, YKL201c] 1e-04
 [BLOCKS] PF00748F
 [BLOCKS] BL01182E Glycosyl hydrolases family 35 proteins
 30 [EC] 2.7.1.37 Protein kinase 7e-06
 [EC] 5.99.1.2 DNA topoisomerase 4e-06
 [PIRKW] phosphotransferase 7e-06
 [PIRKW] pre-mRNA splicing 1e-06
 [PIRKW] citrulline 3e-06
 35 [PIRKW] tandem repeat 3e-06
 [PIRKW] DNA binding 4e-06
 [PIRKW] DNA replication 4e-06
 [PIRKW] isomerase 4e-06
 [PIRKW] ATP 3e-06
 40 [PIRKW] phosphoprotein 1e-06
 [PIRKW] calcium binding 3e-06
 [PIRKW] alternative splicing 7e-06
 [PIRKW] P-loop 3e-06
 [PIRKW] EF hand 3e-06
 45 [PIRKW] hair 3e-06
 [SUPFAM] DEAD/H box helicase homology 3e-06
 [SUPFAM] unassigned Ser/Thr or Tyr-specific protein kinases 4e-
 06
 [SUPFAM] calmodulin repeat homology 3e-06
 50 [SUPFAM] unassigned ribonucleoprotein repeat-containing proteins
 1e-06
 [SUPFAM] unassigned DEAD/H box helicases 3e-06
 [SUPFAM] trichohyalin 3e-06
 [SUPFAM] protein kinase homology 4e-06
 55 [SUPFAM] eukaryotic type I DNA topoisomerase 4e-06
 [SUPFAM] ribonucleoprotein repeat homology 1e-06
 [KW] All_Alpha
 [KW] LOW_COMPLEXITY 22.75 %

5 SEQ ATFSGHGSKMAIPGRQYGLILPKKTQQLHPVLQKPSVFGNDSDDDDDETSVSESLQREAAK
SEGxxxxxxxxxxxxxxxx.....
PRD cccccccccccccccccceeeccccccccccccccccccccccccccccchhhhhhhhhh

10 SEQ KQAMKQTKLEIQKALAEDATVYEYDSIYDEMQKKKEENNPKLLL GKDRPKYIHNLLKAV
SEGxxxxxxxxxxxxxxxx.....
PRD hhhhhhhhhhhhhhhhhhhhhccccccccchhhhhhhhhhhchhhhhhhccccchhhhhhhhhh

15 SEQ LDVTKQKDL SGFYRHLLNQAVGEEV PKCSFREARSGIKEEKSRGFSNEVSSKNRIPQEK
SEG
PRD hhhhhhhccchhhhhhhhhhhhhccccccccccccchhhhhhhhhhhhhhhhhhhhhhhhhhh

20 SEQ CILQTDVKVEENPDADSDFDAKSSADDEIEETRVNCRREKVIETPENDFKHHRSQNHRSRS
SEGxxxxxxxxxxxxxxxx.....
PRD hhhhhhhhhhhhhccchhhccccc

25 SEQ PSEERGHSTRHHTKGSRTSRGHEKREDQHQKQSRDQENHYTDRDYRKERDSHRHREASH
SEG
PRD cccccchhhhhhhhhhhhhhhccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhchh

30 SEQ RDSHWKRHEQEDKPRARDQREERSDRVWKREKDREKYSQREQERDRQNDQNRPSEKGEKE
SEGxxxxxxxxxxxxxxxx.....xxxxxx
PRD hhhhhhhhhcccccchhh

35 SEQ EKSKAKEEHMKVRKERYENNDKYRDREKREVGVSERNQDRKESSPNSRAKDKFLDQER
SEGxxxxxxxx.....
PRD hhhhhhhhhhhhhhhhhhhhhccchhhhhhhhhhhhhhhhhhhchhhhhhhhhhhhhhhhhhhhh

40 SEQ SNKMRNMAKDKERNQEKPSNSESSLGAKHRLTEEGQEKGKEQERPPEAVSKFAKRNEET
SEGxxxxxxxx.....
PRD hhhhhhhhhhhhhhhhhhhhhccchhhhhccchhhhhhhhhhhhhhhhhhhccccchhhhhhhccccc

45 SEQ VMSARDRYLARQMARVNAKTYIEKEDD
SEG
PRD hhhhhhhhhhhhhhhhhhhhhchhhhhccccc

(No Prosite data available for DKFZphtes3_21k14.2)

(No Pfam data available for DKFZphtes3_21k14.2)

DKFZphtes3_22i11

5 group: testis derived

DKFZphtes3_22i11 encodes a novel 580 amino acid protein with similarity to RCC1-like G exchanging factor RLG, UVR8 (UVB-resistance protein) of *Arabidopsis thaliana* and to the murine retinitis pigmentosa GTPase regulator.

No informative BLAST results; No predictive prosite, pfam or SCOP motife.

15 The new protein can find application in studying the expression profile of testis-specific genes.

Homo sapiens chromosome 7q22 sequence, ORF4, extension

20 differences to genmodel of ORF4,
differential splicing

Sequenced by LMU

25 Locus: /map="7q22"

Insert length: 2236 bp

Poly A stretch at pos. 2197, polyadenylation signal at pos. 2180

```

30
    1 ACAATGCTCA GATCGGGAGG TGGAGCCAAT CAGGTCCAAC CAAGAGGAGG
    51 GGACACCGGC ACTCCACTAG CAGGAAAACG GGCCGAGGGA CCGCAAGCAG
   101 GGGGTGCCTA GTCCTCGTCC CCCAAAGACC AATCGTAAGC CAGATACAGG
   35 151 CGAGTGACTG TCAAGAAGGC CAATTAGAGC CTCCGAAGGG AATCTGGACC
    201 TGCCTCTTCT CTGAGGGACG GCTCTACCTA CCAATAGCAT GGGCGAGAAG
    251 GCGGTCCCTT TGCTAAGGAG GAGGCGGGTG AAGAGAAGCT GCCCTTCTTG
    301 TGGCTCGGAG CTTGGGGTTG AAGAGAAGAG GGGGAAAGGA AATCCGATTT
    351 CCATCCAGTT GTTCCCCCA GAGCTGGTGG AGCATATCAT CTCATTCTTC
   40 401 CCAGTCAGAG ACCTTGTTGC CCTCGGCCAG ACCTGCCGCT ACTTCCACGA
    451 AGTGTGCGAT GGGGAAGGCG TGTGGAGACG CATCTGTCGC AGACTCAGTC
    501 CGCGCCTCCA AGATCAGGGT TCTGGAGTCC GGCCCTGGAA GAGAGCTGCC
    551 ATTCTGAAC TACGAAGGG CCTGTATTTC CAGGCATTTC GAGGCCGCCG
    601 CCGATGTCTC AGCAAGAGCG TGGCCCCCTT GCTAGCCCAC GGCTACCGCC
   45 651 GCTTCTTGCC CACCAAGGAT CACGTCTTCA TTCTTGACTA CGTGGGGACC
    701 CTCTTCTTCC TCAAAAATGC CCTGGTCTCC ACCCTCGGCC AGATGCAGTG
    751 GAAGCGGGCC TGTCGCTATG TTGTGTTGTG TCGTGGAGCC AAGGATTTTG
    801 CCTCGGACCC AAGGTGTGAC ACAGTTTACC GTAAATACCT CTACGTCTTG
    851 GCCACTCGGG AGCCGCAGGA AGTGGTGGGT ACCACCAGCA GCCGGGCCTG
   50 901 TGACTGTGTT GAGGTCTATC TGCAGTCTAG TGGGCAGCGG GTCTTCAAGA
    951 TGACATTCCA CCACTCAATG ACCTTCAAGC AGATCGTGCT GGTTGGTCAG
   1001 GAGACCCAGC GGGCTCTACT GCTCCTCACA GAGGAAGGAA AGATCTACTC
   1051 TTTGGTAGTG AATGAGACCC AGCTTGACCA GCCACGCTCC TACACGGTTC
   1101 AGCTGGCCCT GAGGAAGGTG TCCCACTACC TGCCCTACCT GCGCGTGGCC
   55 1151 TGCATGACTT CCAACCAGAG CAGCACCTC TACGTCACAG ACCAGGGGGG
   1201 AGTGTATTTT GAGGTGCATA CCCCAGGGGT GTATCGCGAT CTCTTTGGGA
   1251 CCCTTCAAGC CTTTGACCCC CTGGACCAGC AGATGCCGCT TGCTCTCTCA
   1301 CTGCCTGCCA AGATCCTATT CTGTGCTCTT GGCTACAACC ACCTTGGCCT

```

1351 GGTGGATGAA TTTGGCCGAA TCTTCATGCA AGGAAATAAC AGATACGGGC
 1401 AGCTAGGAAC AGGGGACAAA ATGGACCGAG GGGAAACCCAC ACAGGTTTGT
 1451 TACCTGCAGC GGCCCATCAC CCTGTGGTGC GGCCTCAACC ACTCCCTGGT
 1501 GCTGAGCCAG AGCTCAGAGT TCAGCAAGGA GCTGCTGGGC TCGGGCTGTG
 5 1551 GGGCTGGGGG CCGCCTCCCA GGCTGGCCCA AGGGGAGTGC CTCCTTCGTC
 1601 AAGCTCCAAG TCAAGGTCCC TCTGTGTGCC TGTGCCCTCT GTGCCACCAG
 1651 GGAGTGCCTA TACATCCTGT CCAGCCACGA CATTGAGCAG CACGCCCCCT
 1701 ATCGCCACCT GCCAGCCAGC AGGGTGGTGG GGA CTCTCTGA GCCCAGCCTG
 1751 GGGGCCAGAG CACCCACAGG CCCCAGGGGG ATGGCCACAG CCTGCGAGGA
 10 1801 GTACCTCAGC CAGATCCACA GTTGCCAAAC GTTGCAGGAC CGCACGGAGA
 1851 AGATGAAGGA GATCGTAGGG TGGATGCCCC TGATGGCCGC ACAGAAGGAC
 1901 TTCTTCTGGG AGGCCCTGGA CATGCTGCAG AGGGCTGAAG GAGGCGGGGG
 1951 TGGTGTAGGG CCCCCAGCCC CTGAGACCTA ATCCCCCTCA TGCTAGCCTA
 2001 GTCCCTGGAG GAGGGAGTCC GGCCCCAGGC CAGGGACTAA GGAGCAATGA
 15 2051 CCATTGTGCA CATGCGTGTG GGAAGGGGTT GCTAGGGGGT GGGGACGGCT
 2101 AACCAGGGTA AGAATGTTCA GGGGGCTGCC CAGGAGGGGC CCCCACCTG
 2151 ACTATCATGG ACAAGAGATT TGATGGATAG AATAAAAGGC TGCAGCGAAA
 2201 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAG

20

BLAST Results

Entry AF053356 from database EMBL:

25 Homo sapiens chromosome 7q22 sequence, complete sequence.

Score = 2952, P = 0.0e+00, identities = 666/729

10 exons

30

Medline entries

No Medline entry

35

Peptide information for frame 2

40

ORF from 239 bp to 1978 bp; peptide length: 580

Category: similarity to unknown protein

Classification: no clue

45 1 MGEKAVPLLR RRRVKRSCPS CGSELGVEEK RGKGNPISIQ LFPPPELVEHI
 51 ISFLPVRDLV ALGQTCRYFH EVCDGEGVWR RICRRRLSPRL QDQSGVVRPW
 101 KRAAILNYTK GLYFQAFGGR RRCLSKSVAP LLAHGYRRFL PTKDHFVILD
 151 YVGTLFFLKN ALVSTLGQMQ WKACRYVVL CRGAKDFASD PRCDTVYRKY
 201 LYVLATREPQ EVVGTTSSRA CDCVEVYLQS SGQRVFKMTF HHSMTFKQIV
 50 251 LVGQETQRAL LLLTEEGKIY SLVVNETQLD QPRSYTVQLA LRKVSHYLP
 301 LRVACMTSNQ SSTLYVTDAQ GYVFEVHTPG VYRDLFGTLQ AFDPLDQQMP
 351 LALSIPAKIL FCALGYNHLG LVDEFGRIFM QGNNRYGQLG TGDKMDRGE
 401 TQVCYLQRP TQVCYLQRP TQVCYLQRP TQVCYLQRP TQVCYLQRP
 451 ASFKVLQVKV PLCACALCAT RECLYLSSH DIEQHAPYRH LPASRVVGT
 55 501 EPSLGARAPQ DPGGMAQACE EYLSQIHSCQ TLQDRTEKMK EIVGWMLPMA
 551 AQKDFWEAL DMLQRAEGGG GGVGPPAPET

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_22i11, frame 2

TREMBL:AF053356_11 product: "ORF4"; Homo sapiens chromosome 7q22 sequence, complete sequence., N = 1, Score = 1554, P = 1.6e-159

TREMBL:AF130441_1 gene: "UVR8"; product: "UVB-resistance protein UVR8"; Arabidopsis thaliana UVB-resistance protein UVR8 (UVR8) mRNA, complete

cds., N = 1, Score = 109, P = 0.0082

TREMBL:AF044677_1 gene: "Rpgr"; product: "retinitis pigmentosa GTPase regulator"; Mus musculus retinitis pigmentosa GTPase regulator (Rpgr) mRNA, complete cds., N = 1, Score = 106, P = 0.035

>TREMBL:AF053356_11 product: "ORF4"; Homo sapiens chromosome 7q22 sequence, complete sequence.
Length = 318

HSPs:

Score = 1554 (233.2 bits), Expect = 1.6e-159, P = 1.6e-159
Identities = 303/318 (95%), Positives = 303/318 (95%)

Query: 1
MGEKAVPLLRRRRVKRSCPCSGSELGVEEKRKGKGNPISIQLFPPPELVEHIISFLPVRDLV 60

MGEKAVPLLRRRRVKRSCPCSGSELGVEEKRKGKGNPISIQLFPPPELVEHIISFLPVRDLV
Sbjct: 1
MGEKAVPLLRRRRVKRSCPCSGSELGVEEKRKGKGNPISIQLFPPPELVEHIISFLPVRDLV 60

Query: 61
ALGQTCRYFHEVCDGEGVWRRICRRLSPRLQDQGSQVWPWKRAAILNYTKGLYFQAFGGR 120
ALGQTCRYFHEVCDGEGVWRRICRRLSPRLQDQ

TKGLYFQAFGGR
Sbjct: 61 ALGQTCRYFHEVCDGEGVWRRICRRLSPRLQDQD-----
TKGLYFQAFGGR 106

Query: 121
RRCLSKSVAPLLAHGYRRFLPTKDHVFILDYVGTLFFLKNALVSTLGQMQRACRYVVL 180

RRCLSKSVAPLLAHGYRRFLPTKDHVFILDYVGTLFFLKNALVSTLGQMQRACRYVVL
Sbjct: 107
RRCLSKSVAPLLAHGYRRFLPTKDHVFILDYVGTLFFLKNALVSTLGQMQRACRYVVL 166

Query: 181
CRGAKDFASDPRCDTVYRKLYVLATREPQEVVGTSSRACDCVEVYLQSSGQRVFKMTF 240

CRGAKDFASDPRCDTVYRKLYVLATREPQEVVGTSSRACDCVEVYLQSSGQRVFKMTF


```
SEQ  FCALGYNHLGLVDEFGRIFMQGNNRYGQLGTGDKMDRGEPTQVCYLQRPITLWCGLNHSL
SEG  .....
PRD  eeeeeccccceeeeeceeeeeccccccccccccccccccccccccccccccccccccceeeeeccccceeeeeccceee

5  SEQ  VLSQSSSEFSKELLGCGCGAGGRLPGWPKGSASFVKLQVKVPLCACALCATRECLYLSSH
SEG  .....xxxxxxxxxxxxxxxx.....
PRD  eeeeeccccceeeeeccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccceee

10  SEQ  DIEQHAPYRHLPASRVVGTPEPSLGARAPQDPGGMAQACEEYLSQIHSCQTLQDRTEKMK
SEG  .....
PRD  cccccccccccccccccceccccccccccccccccccccccccchhhhhhhhhhhhhcchhhhhhhhhhh

SEQ  EIVGWMLMAAQKDFFWREALDMLQRAEGGGGGVGPPAPET
SEG  .....xxxxxxx.....
15  PRD  hhhhcchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccccceeecccccc
```

(No Prosite data available for DKFZphtes3_22i11.2)

20 (No Pfam data available for DKFZphtes3_22i11.2)

5 group: testis derived

DKFZphtes3_22124 encodes a novel 451 amino acid protein with similarity to the F-box protein FBL2 of the rat.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motif.

The new protein can find application in studying the expression profile of testis-specific genes.

15

similarity to p37NB (Homo sapiens)

Sequenced by LMU

20

Locus: /map="7q22-q31.1"

Insert length: 1537 bp

Poly A stretch at pos. 1459, no polyadenylation signal found

25

```

    1 CAACAGGACG ATGCGACTCC TGCCGAGGCA CTTCCACAAC TTACAGAATC
   51 TTAGTTTGGC TTATTGCAGA CGGTTACAG ACAAAGGCTT ACAGTACCTG
  101 AACTTGGGGA ATGGATGCCA CAAGCTCATC TATCTGGACC TCTCTGGCTG
  151 CACCCAGATT TCAGTCCAAG GCTTCAGGTA CATTGCAAAAC AGCTGCACTG
  201 GAATTATGCA TCTTACCATT AATGACATGC CAACTCTGAC GGACAACCTGT
  251 GTAAAAGCTT TAGTTGAAAA ATGCTCTCGT ATTACATCGC TGGTTTTTAC
  301 TGGTGACCCG CATATCTCCG ATTGTACTTT CAGAGCTCTT TCTGCTTGTA
  351 AACTCAGAAA GATCCGATTT GAAGGAAATA AAAGGGTTAC TGATGCATCC
  401 TTCAAATTTA TAGACAAGAA TTATCCAAAT CTCAGTCACA TTTATATGGC
  451 TGACTGCAAG GGAATAACAG ACAGCAGCCT CAGATCCCTT TCACCTTTGA
  501 AGCAACTGAC TGTGTTGAAT TTGGCAAATT GTGTAAGAAT TGGTGATATG
  551 GGACTAAAGC AATTTCTTGA TGGTCCTGCA AGCATGAGGA TAAGAGAGCT
  601 AAATTTAAGC AACTGTGTGC GGCTAAGTGA TGCCTTTGTT ATGAAACTAT
  651 CTGAGCGCTG CCCTAATTTA AACTACTTGA GTTTACGAAA TTGTGAACAT
  701 TTGACTGCCC AAGGAATTGG ATATATTGTA AACATCTTTT CTTGGGTATC
  751 AATAGATCTC TCTGGAACAG ACATCTCTAA TGAGGGTTTG AATGTGCTTT
  801 CCAGACATAA AAAATTGAAG GAACCTTCTG TATCTGAATG TTATAGAATC
  851 ACTGATGATG GAATTCAGGC ATTCTGCAAA AGCTCACTGA TCTTGGAAAC
  901 TTTGGATGTC TCTTATTGCT CCCAGCTGTC AGATATGATT ATCAAAGCAC
  951 TGGCCATTTA CTGCATTAAC CTCACATCTC TCAGCATTGC TGGCTGTCCA
 1001 AAGATTACTG ACTCAGCAAT GGAGATGTTA TCGGCAAAAT GCCATTACCT
 1051 GCACATTTTG GATATCTCTG GTTGTGTCTT GCTTACTGAC CAAATCCTTG
 1101 AGGACCTTCA GATAGGCTGC AAACAACCTC GGATCCTTAA GATGCAATAC
 1151 TGCACAAATA TTTCCAAGAA GGCAGCTCAA AGAATGTCAT CTAAAGTTCA
 1201 GCAGCAGGAA TACAACACTA ATGACCCTCC ACGTTGGTTT GGCTATGATA
 1251 GCGAAGGAAA CCTGTGTACA GAGCTTGACA ACATAACATC ATCTAAAGGA
 1301 GCCTTAGAAT TAACAGTGAA AAAGTCAACA TACAGCAGTG AAGACCAAGC
 1351 AGCGTGACCT TCAGCCTCAA GCAGGAAGAA CAAAAAATCA AGAAGTTGGC
 1401 AAGTTTTCTC CATTTGTTGC AAGTATGTTT ACTAGCTGAA TCTCAATAAC
 1451 AATGTAAACA AGCAAAAAAA AAAAAAATAA AAAAAAATAA AAAAAAATAA
 1501 AAAAAAATAA AAAAAAATAA AAAAAAATAA AAAAAAATAA AAAAAAATAA
```

BLAST Results

- 5 Entry AC005250 from database EMBL:
Homo sapiens BAC clone RG318M05 from 7q22-q31.1, complete
sequence.
Score = 830, P = 1.8e-124, identities = 180/193
- 10 Entry HS32907 from database EMBL:
Human p37NB mRNA, complete cds.
Score = 318, P = 4.6e-04, identities = 70/78

15

Medline entries

- 97136875:
20 Kim D, LaQuaglia MP, Yang SY.; A cDNA encoding a putative 37 kDa
leucine-rich repeat
(LRR) protein, p37NB, isolated from S-type neuroblastoma
cell has a differential tissue distribution. Biochim Biophys Acta
1996 . .
25 Dec 11;1309(3):183-8

30

Peptide information for frame 2

- ORF from 11 bp to 1354 bp; peptide length: 448
Category: similarity to known protein
35 Classification: unclassified

1 MRLLPRLFHN LQNLSLAYCR RFTDKGLQYL NLGNGCHKLI YLDLSGCTQI
51 SVQGFYRIAN SCTGIMHLTI NDMPTLTDNC VKALVEKCSR ITSLVFTGAP
101 HISDCTFRAL SACKLRKIRF EGNKRVT DAS FKFDKNYPN LSHIYMADCK
40 151 GITDSSLRSL SPLKQLTVLN LANCVRIGDM GLKQFLDGPA SMRIEELNLS
201 NCVRLSDAFV MKLSERCPNL NYLSLRNCEH LTAQGIGYIV NIFSLVSI DL
251 SGTDISNEGL NVLSRHKKLK ELSVSECYRI TDDGIQAFCK SSLILEHL DV
301 SYCSQLSDMI IKALAIYCIN LTSLSIAGCP KITDSAMEML SAKCHYLHIL
351 DISGCVLLTD QILEDLQIGC KQLRILKMQY CTNISKKAAQ RMSSKVQQQE
45 401 YNTNDPPRWF GYDREGNPVT ELDNITSSKG ALELTVKKST YSSEDQAA

BLASTP hits

50

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_22124, frame 2

55 No Alert BLASTP hits found

Pedant information for DKFZphtes3_22124, frame 2

Report for DKFZphtes3_22124.2

```

5  [LENGTH]  451
   [MW]      50545.95
   [pI]      8.68
   [HOMOL]   TREMBLNEW:AF186273_1 product: "leucine-rich
10  repeats containing F-box protein FBL3"; Homo sapiens leucine-rich
   repeats containing F-box protein FBL3 mRNA, complete cds. 8e-31
   [FUNCAT]  11.01 stress response [S. cerevisiae, YJR090c] 8e-20
   [FUNCAT]  03.01 cell growth [S. cerevisiae, YJR090c] 8e-20
   [FUNCAT]  08.19 cellular import [S. cerevisiae, YJR090c] 8e-20
   [FUNCAT]  03.22 cell cycle control and mitosis [S. cerevisiae,
15  YJR090c] 8e-20
   [FUNCAT]  03.04 budding, cell polarity and filament formation
       [S. cerevisiae, YJR090c] 8e-20
   [FUNCAT]  01.05.04 regulation of carbohydrate utilization [S.
20  cerevisiae, YJR090c] 8e-20
   [FUNCAT]  11.04 dna repair (direct repair, base excision repair
       and nucleotide excision repair) [S. cerevisiae, YJR052w] 3e-07
   [FUNCAT]  30.10 nuclear organization [S. cerevisiae, YJR052w]
       3e-07
   [BLOCKS]  PRO0019B
25  [BLOCKS]  PRO0364D
   [BLOCKS]  BP01721A
   [BLOCKS]  BP03743B
   [PIRKW]   tandem repeat 2e-18
   [PIRKW]   zinc finger 1e-07
30  [PIRKW]   DNA binding 1e-07
   [SUPFAM]  leucine-rich alpha-2-glycoprotein repeat homology 2e-18
   [SUPFAM]  regulatory protein ESAG8c 1e-07
   [KW]      Alpha_Beta
35

   SEQ  NRTMRLLP RHFHNLQNL SLAYCRRFTDKGLQYLN LGNGCHKLIYLDLSGCTQISVQGFY
   PRD  cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

40  SEQ  IANSCTGIMHLTINDMPTLT DNCVKALVEKCSRITSLVFTGAPHISDCTFRALSACKLRK
   PRD  cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

   SEQ  IRFEGNKRVTDASFKFIDKNYPNL SHIYMADCKGITDSSLRSLSPKQLTVLNLANCVRIR
   PRD  eccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

45  SEQ  GDMGLKQFLDGPASMRIRELNLSN CVRLSDAFVMKLSERCPNLNYLSLRNCEHLTAQGIG
   PRD  cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

   SEQ  YIVNIFSLVSLDLSGTDISNEGLNVLSRHKKLKELSVSECYRITDDGIQAFCKSSLILEH
50  PRD  eccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

   SEQ  LDVSYCSQLSDMIKALAIYCINLTSL SIAGCPKITDSAMEMLSAKCHYLHILDISGCVL
   PRD  cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

55  SEQ  LTDQILEDLQIGCKQLRILKMQYCTNISK KAAQRMSSKVQQAQ EYNTNDPPRWFGYDREGN
   PRD  chhhhhhhhhhhcchhhhhhhc cccccchhhhhhhhhhhhhhecccccccccccccccccc

   SEQ  PVTELDNITSSKGALELTVKKSTYSS EDQAA

```

PRD ccccccccccccccccccccccccccccccccc

- 5 (No Prosite data available for DKFZphtes3_22124.2)
(No Pfam data available for DKFZphtes3_22124.2)

DKFZphtes3_26g3

5 group: testis derived

DKFZphtes3_26g3 encodes a novel 1090 amino acid protein without similarity to known proteins.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motife.

The new protein can find application in studying the expression profile of testis-specific genes.

15

similarity to C.elegans CD9D4.4

on genomic level encoded by HSDJ198I9
20 perhaps complete cds.

Sequenced by EMBL

Locus: /map="b"

25

Insert length: 4562 bp

Poly A stretch at pos. 4550, polyadenylation signal at pos. 4515

30 1 GATTCAGTTA CTGAAGACTT AGATGCACCC TGGATGGGAA TTCAGAATCT
51 TCAGAGATCA GAGTCCAGTA AAATGGATAA ATATGAGACT GAAGAAAGCT
101 CTGTAGCAGG ACTTTCTAGC CCAGAGTTGA AAGTCAGACC TGCTGGTGCC
151 TCCAGTATTT GGTATACAGA AGGTGAAAAG CAGCTAACAA AATCTCTAAA
201 AGGAAAGAAT GAAGAATCAA ATAAATCCAA AGTTAAGGTT ACTAAGCTTA
35 251 TGAAAACAAT GAAATCTGAA AACACAAAAA AATTAATAAA ACAGAACTCT
301 AAGGATTCTG TGGTTTTGGT AGGCTACAAA TGTTTGAAAA GTACAGCATC
351 AAATGATCTC ATTAAATGCT TTGAAGGCCAA TCCTTCACAT AGTCAGAAGG
401 AAGGTCTGGA TCCCACAATA TGTGGATATA ATTTTGACCC AAAGACCTAC
451 ATGAGACAGA CAAGTCAAAA GGAAGCTAGC TGTTTGCCAA CTAATACAGA
40 501 GAGAACTGAA CAAAAGTCTC CAGATATTGA AAATGTTCAA CCAGACCAGT
551 TTGATCCTTT GAACTCTGGC AACCTAAATC TTTGTGCAAA TTTGTCCATT
601 TCAGGTAAAC TTGATATCTC CCAGGACGAT AGTGAAATTA CACAAATGGA
651 ACACAATCTG GCATCCAGAA GGTCAATCAGA CGATTGCCAT GATCATCAAA
701 CAACCCCATC TTTGGGAGTT AGAACAATTG AAATAAAGCC CAGTAATAAA
45 751 GATCCTTTCA GTGGAGAGAA TATAACTGTC AAATAAGGAC CTTGGACAGA
801 GCTTCGACAA GAGGAAATAC TTGTGGATAA TTTACTACCC AACTTTGAGT
851 CCTTAGAATC TAATGGTAAA TCTAAATCTA TAGAAATAAC ATTTGAAAAG
901 GAAGCTTTGC AAGAAGCAAA GTGTCTTTCT ATTGGAGAAT CATTAATAA
951 ATTACGAAGT AATCTACCTG CCCCTTCTAC AAAAGAATAT CATGTTGTAG
50 1001 TAAGTGGAGA TACAATTAAG TTACCAGATA TTAGTGCCAC ATATGCCTCA
1051 TCTAGATTTT CAGATTCCAGG TGTTGAAAAGT GAACCGAGTT CTTTTGCGAC
1101 ACATCCAAAC ACTGATTTAG TCTTTGAAAC TGTGCAAGGG CAAGGTCCTT
1151 GCAATAGTGA AAGATTATTT CCTCAGCTTT TGATGAAACC TGATTATAAT
1201 GTAAAATTTT CATTAGGAAA TCATTGTACT GAGAGTACAA GTGCTATAAG
55 1251 TGAAATACAG TCATCTTTGA CATCCATAAA CTCTCTACCC TCCGATGATG
1301 AACTGTCACC TGATGAAAAT TCTAAGAAAT CTGTTGTACC TGAATGCAT
1351 CTAAATGATA GCAAACTGT ATTAATCTA GGAACGACTG ATTTGCCAAA
1401 ATGTGATGAT ACTAAAAAGT CAAGTATCAC TTTGCAACAG CAGAGTGTG

	1451	TATTTTCAGG	GAAC TTGGAC	AATGAAACTG	TAGCAATACA	TTCTTAAAT
	1501	TCAAGCATT	AAGACCCTTT	ACAATTTGTT	TTTTTCAGATG	AAGAGACTTC
	1551	CAGTGATGTG	AAAAGTAGTT	GCAGCTCCAA	ACCTAACTTG	GATACTATGT
	1601	GTAAGGCTT	CCAGAGTCCT	GATAAATCTA	ATAACTCTAC	AGGGACAGCA
5	1651	ATTACATTAA	ATTCAAAAC	GATTTGTTTA	GGCACTCCTT	GTGTCATTTC
	1701	AGGTTCATT	TCTAGTAATA	CAGATGTTAG	TGAAGATAGA	ACTATGAAAA
	1751	AAAATAGTGA	TGTATTAAAT	CTCACACAGA	TGTATTGAGA	AATCCCTACA
	1801	GTTGAAAGTG	AAACTCATCT	GGGTACAAGT	GATCCTTTTT	CAGCCAGTAC
	1851	TGATATAGTA	AAGCAAGGGC	TTGTGGAAAA	TTATTTTGGT	TCTCAAAGCA
10	1901	GTACGGATAT	TTCTGACACA	TGTGCTGTTA	GCTACAGCAA	TGCACTTAGC
	1951	CCTCAGAAGG	AAACTTCTGA	AAAAGAAATT	AGTAATCTTC	AGCAGGAACA
	2001	GGATAAAGAG	GATGAGGAGG	AAGAGCAGGA	TCAACAAATG	GTTCAAAATG
	2051	GGTACTATGA	AGAAACAGAT	TATTCAGCTT	TGGATGGAAC	AATAAATGCT
	2101	CACATATACAA	GCAGAGATGA	ACTAATGGAA	GAAAGACTTA	CAAAATCTGA
15	2151	AAAAATAAAC	AGTGAATATC	TGAGAGATGG	TATAAACATG	CCTACTGTCT
	2201	GTACTTCTGG	TTGTTTGTCC	TTCCCGTCTG	CACCACGAGA	GTCTCCTTGT
	2251	AATGTAAAT	ATTCTTCCAA	AAGTAAATTT	GATGCCATTA	CAAAGCAGCC
	2301	AAGCAGTACT	TCTTACAAC	TCACTTCTTC	GATTTCTCTG	TATGAAAGTT
	2351	CACCAAAACC	TCAAATACAA	GCCTTCCTTC	AGGCAAAAGA	AGAAGTGAAG
20	2401	CTACTAAAAC	TTCTTGGGTT	CATGTACAGT	GAAGTTCCTC	TGCTGGCATC
	2451	CTCAGTACCT	TATTTTAGTG	TAGAAGAAGA	GGGTGGTTCT	GAAGATGGAG
	2501	TACATCTGAT	TGCTGTGTG	CACGGTTTAG	ATGGAAACAG	TGCAGATCTC
	2551	CGATTCTGTA	AAACTTACAT	TGAACCTGGA	TTGCCTGGGG	GAGAATTTGA
	2601	TTTTCTTATG	TCTGAGAGAA	ATCAGAATGA	TACTTTTGCT	GATTTTGATA
25	2651	GCATGACTGA	TCGTCTTTTG	GATGAGATAA	TACAGTATAT	TCAGATATAT
	2701	AGTCTAACAG	TCTCAAAAAT	AAGCTTTATT	GGACATTCTG	TGGGCAATTT
	2751	AATAATTCTG	TCAGTGCTTA	CAAGGCCAAG	GTTTAAATAT	TACCTCAACA
	2801	AACTTCATAC	CTTCTGTCT	CTTCTGGAC	CTCACCTTGG	TACACTCTAC
	2851	AACAGCAGTG	CTCTTGTTAA	TACAGGTCTC	TGGTTTATGC	AGAAATGGAA
30	2901	AAAATCAGGT	TCGCTTTTGC	AGCTGACATG	TCGAGATCAC	TCAGACCTTC
	2951	GCCAAACTTT	TTTATATAAG	CTTAGTAACA	AAGCAGGGCT	TCATTATTTC
	3001	AAAAATGTTG	TGCTAGTGGG	ATCCCTACAG	GATCGCTATG	TTCTTTATCA
	3051	CTCTGCCCGC	ATTGAAATGT	GTAAACACAG	TTTAAAGGAC	AAACAGTCAG
	3101	GACAGATCTA	TTTCAAAATG	ATCCACAAC	TGCTTCGACC	CGTTCTGCAA
35	3151	AGCAAGGACT	GTAATTTGGT	TCGCTATAAT	GTCATCAATG	CATTGCCCAA
	3201	TACAGCTGAT	TCACTCATTG	GGAGAGCTGC	ACATATAGCT	GTTCTTGATT
	3251	CGGAAATATT	TTTAGAGAAA	TTCTTTCTGG	TTGCTGCCCT	CAATATTTTC
	3301	CAGATAGTATA	AAAGCATTGT	TTCGCACTGG	ACAATTACCT	CATTCAACAA
	3351	TGTTTCAAAT	AATGTATTAT	ATTAAAATGT	AGATGCTGAT	AAGTTCTAAG
40	3401	AAATATTTAT	ACCTTTTTAT	ATGGAAGATA	ATTTATATCA	TCCATGTTTA
	3451	GTGCTTTTTA	AACATCAACT	TTACTTTCTA	GGTAATGTGG	CTGTGCAATA
	3501	TTTTTTTAAAT	TTTATCTTTT	TACTTTTCTA	TTACTTTTTT	ATATATTTTG
	3551	CTACCTAAGT	ATTTCAAGTGA	AACTTTAAGC	CCATACCTGT	GTCTGATTGT
	3601	TTATTATTGG	CTTTCACAAA	TTCTTACATC	AGACTACATT	ATATTAGAGA
45	3651	CCATTATTGC	TAGAATAGCA	TGGGATTTAA	AATTTCTAAT	ACTGGGGGTA
	3701	TTATTTAGTT	AATTATAAAT	TTTTCTTTTC	ACATTTTACT	GTGTTTTAAC
	3751	TGGAAATAAA	ATTATGGCTG	CTACAATATA	TTTTTTGAAA	TCAACTTCTG
	3801	TAGTTCTAAA	ATACAACCTT	ATCATACAAT	CAAACCAGGT	AGTTTATATA
	3851	AAACAGTGTA	ATACAAGTTT	TCTATAAAGT	CATTACTGTT	GCTTAAACAT
50	3901	ATTTTCATGCC	TATTTAAATA	TATTTTCTAC	TGGTGATTTT	AACATTATTT
	3951	CTCATACTGA	CTTTTATTAC	TGGAAATGTT	CCTGTACATG	TTGGCAGCAG
	4001	ATAAAGATTT	TTGAATGTTT	GAATGCCCTC	TGCCCTTGATT	TGGTTGGATT
	4051	TTGCTAATTG	GTATGTTGCT	TGAACCTTAT	GACTACATTT	TCTTTTAACT
	4101	TTTTTCATGG	ACTTCCTTAT	ATGTACATAA	TAATTAATG	TTGAAATTTA
55	4151	TGAAATACTT	TTATGAATTT	AGATAATTTT	TAAATATTGT	TAAAATTTAT
	4201	TGAACATAAA	AGTAATGTAA	ATAAAATAAT	TCATGTTAAA	GATGGAACAA
	4251	AATAATTAAAC	TTTACATGTT	TGGTGATACA	GATGCAAATG	TTTTTGATAT
	4301	ATGGAGATGT	TGAGTCTTTT	GACTTTACTA	AAGGTGCTGA	ATAGCATTA

4351 ATTCACTATT TTCCTTTTCT GTTTTACTTG TGAAAATAAA AATGCACTAA
4401 GGTGGGTAG AAGTTCTGTT TGCACCTACT AATTGTGACA GACAGAGGTT
4451 TTTGTAAGTA TTTATTGTAC AATTGATGCA TGTTTATTTT TAGCGTTGTT
4501 ATTGCTCTG GTGTTAATAA ATGAACAAAT GGCTATCTGG AGGAACAGCT
5 4551 AAAAAAAAAA AA

BLAST Results

10 Entry HSDJ198I9 from database EMBLNEW:
Human DNA sequence *** SEQUENCING IN PROGRESS *** from clone
DJ198I9
15 Score = 7221, P = 0.0e+00, identities = 1455/1461

Medline entries

20 No Medline entry

Peptide information for frame 1

25 ORF from 34 bp to 3303 bp; peptide length: 1090
Category: similarity to unknown protein
30 Classification: no clue

1 MGIQNLQRSE SSKMDKYETE ESSVAGLSSP ELKVRPAGAS SIWYTEGEKQ
51 LTKSLKGKNE ESNKSKVKVT KLMKTMKSEN TKKLIKQNSK DSVVLVGYKC
101 LKSTASNDLI KCFEGNPSHS QKEGLDPTIC GYNFDPKTYM RQTSQKEASC
35 151 LPTNTERTEQ KSPDIENVQP DQFDPLNSGN LNLCANLSIS GKLDISQDDDS
201 EITQMEHNLA SRRSSDDCHD HQTTPSLGVR TIEIKPSNKD PFSGENITVK
251 LGPWTELREQ EILVDNLLPN FESLESNGKS KSIEITFEKE ALQEAKCLSI
301 GESLTKLRSN LPAPSTKEYH VVVSQDTIKL PDISATYASS RFSDSGVESE
351 PSSFATHPNT DLVFETVQGG GPCNSERLFP QLLMKPDYNV KFSLGNHCTE
40 401 STSAISEIQS SLTSINSLPS DDELSPDENS KKSVPPECHL NDSKTVLNLG
451 TTDLPKCDDT KKSSITLQQQ SVVFSGNLDN ETVAIHSLNS SIKDPLQFVF
501 SDEETSSDVK SSCSSKPNLD TMCKGFQSPD KSNNSTGTAI TLNSKLICLG
551 TPCVISGSIS SNTDVSEDRT MKKNSDVLNL TQMYSEIPTV ESETHLGTSD
601 PFSASTDIVK QGLVENYFGS QSSTDISDTC AVSYSNALSP QKETSEKEIS
45 651 NLQEQQDKED EEEEQDQQMV QNGYEEETDY SALDGTINAH YTSRDELMEE
701 RLTKSEKINS DYLRDGINMP TVCTSGCLSF PSAPRESPCN VKYSSKSKFD
751 AITKQPSSTS YNFTSSISWY ESSPKPQIQ A FLQAKEELKL LKLPGFMYSE
801 VPLLASSVPY FSVEEEGGSE DGVHLIVCVH GLDGNSADLR LVKTYIELGL
851 PGGRIDFLMS ERNQNDTFAD FDSMTDRLLD EIIQYIQIYS LTVSKISFIG
50 901 HSLGNLIIRS VLTRPRFKYY LNKLTFLSL SGPHLGTLYN SSALVNTGLW
951 FMQWKKSGS LLQLTCRDHS DPRQTFLYKL SNKAGLHYFK NVVLVGSLQD
1001 RYVPYHSARI EMCKTALKDK QSGQIYSEMI HNLLRPVLQS KDCNLVRYNV
1051 INALPNTADS LIGRAAHIAV LDSEIFLEKF FLVAALKYFQ

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_26g3, frame 1

5 No Alert BLASTP hits found

Pedant information for DKFZphtes3_26g3, frame 1

10

Report for DKFZphtes3_26g3.1

[LENGTH] 1101
 [MW] 122245.22
 15 [pI] 5.12
 [HOMOL] TREMBL:CEAF219b_1 gene: "C09D4.4"; Caenorhabditis
 elegans cosmid C09D4. 2e-38
 [FUNCAT] 99 unclassified proteins [S. cerevisiae, Y0R059c]
 2e-06
 20 [BLOCKS] BLO01208
 [KW] Alpha_Beta
 [KW] LOW_COMPLEXITY 6.72 %

25 SEQ DSVTEDLDAPWMGIQNLQRSESSKMDKYETEESVAGLSSPELKVRPAGASSIWYTEGEK
 SEG
 PRD cccccccccccccccccchhhhhhhhhhhcc

30 SEQ QLTSLKLGKNEESNKSQKVKVTCLMKTMKSENTKKLIKQNSKDSVVLVGYKCLKSTASNDL
 SEGxx
 PRD hhhhhhcc

35 SEQ IKCFEGNPSHSQKEGLDPTICGYNFDPKTYMRQTSQKEASCLPTNTERTEQKSPDIENVQ
 SEG
 PRD eeeeecc

40 SEQ PDQFDPLNSGNLNLCANLSISGKLDISQDDSEITQMEHNLASRRSSDDCHDHQTTPLSGV
 SEG
 PRD ccc

45 SEQ RTIEIKPSNKDPFSGENITVKLGPTWELRQEEILVDNLLPNFESLESNGKSKSIEITFEK
 SEG
 PRD eeeeecc

50 SEQ EALQEAQKCLSIGESLTKLRNLAPSTKEYHVVVSGDTIKLPDISATYASSRFSDSGVES
 SEG
 PRD hhh

55 SEQ EPSSFATHPNTDLVFETVQGGGPCNSERLFPQLLMKPDYNVKFSLGNHCTESTSAISEIQ
 SEG
 PRD ccc

SEQ SSLTSINSLPSDDELSPDENSSKSVVPECHLNDSKTVLNLGTTDLPKCDDTKKSSITLQD
 SEG
 PRD ccc

SEQ QSVVFSGNLDNETVAIHSLNSSIKDPLQFVFSDEETSSDVKSSCSSKPNLDTMCKGFQSP
 SEGxx

(No Pfam data available for DKFZphtes3_26q3.1)

DKFZphtes3_29f24

5 group: signal transduction

DKFZphtes3_29f24 encodes a novel 526 amino acid protein with similarity to murine netla.

10 The closely related mNET1 activates signalling pathways in addition to those directly controlled by activated RhoA. The novel protein is expressed ubiquitously.

15 The new protein can find application in modulation/blocking signalling pathways.

similarity to netla (Mus musculus)

20 perhaps complete cds.

Sequenced by BMFZ

Locus: /map="72.40 cR from top of Chr3 linkage group"

25

Insert length: 3559 bp

Poly A stretch at pos. 3534, polyadenylation signal at pos. 3513

```

30      1 CGCCGCCGCC CGGCATCGTG GAGCTGGGGC CCCCTTTTGC CTGGGAGTTT
      51 TGTAGTCGCC TAGGGTCAGC GGTGACATCC CAAAGGGCAG GCCCGGCAGC
     101 CGCCATGGTG GCCAAGGATT ACCCCTTCTA CCTCACGGTC AAGAGAGCGA
     151 ACTGCAGCCT GGAGCTACCC CCGGCCAGCG GTCCGGCCAA GGACGCTGAG
     201 GAGCCTAGTA ATAAACGGGT CAAACCCCTT TCCCGAGTCA CGTCGCTAGC
     35 251 AAAGCTCATC CCGCCCGTGA AGGCCACGEC ATTAAAGCGC TTCAGTCAAA
     301 CCCTGCAGCG CTCCATTAGC TTCCGCAGTG AGAGCCGCCC TGACATCCTC
     351 GCGCCCCGAC CCTGGTCCAG AAATGCCGCC CCCTCGAGCA CGAAACGGAG
     401 AGATAGCAAG CTGTGGAGTG AGACCTTCGA TGTGTGCGTC AATCAGATGC
     451 TTACATCCAA GGAAATCAAA CGTCAGGAGG CGATCTTTGA GCTTTCCCAA
     501 GGAGAAGAAG ACTTGATAGA AGACTTGAAA TTAGCAAAAA AGGCCTATCA
     551 TGACCCCATG CTGAAACTCT CCATAATGAC AGAACAAGAG TTGAATCAAA
     601 TTTTGGGAAC ACTGGACTCT CTAATTCCTC TACATGAAGA GCTCCTTAGT
     651 CAGCTTCGAG ATGTTAGGAA GCCTGATGGC TCGACTGAAC ATGTTGGTCC
     701 CATCCTCGTG GGCTGGCTCC CTTGCCTCAG CTCCTATGAT AGCTACTGCA
     45 751 GCAATCAAGT AGCCGCCAAA GCTCTGCTGG ACCACAAAAA GCAAGATCAC
     801 CGAGTCCAGG ATTTCCCTACA GCGATGTTTA GAATCCCCCT TTAGCCGCAA
     851 ACTAGATCTC TGGAATTTCC TCGATATTCC AAGAAGCCGC CTGGTAAAAT
     901 ACCCTCTGCT TCTCCGAGAA ATCTTGAGGC ACACACCAAA TGATAATCCA
     951 GATCAGCAGC ACTTGGAAGA AGCTATAAAT ATCATTGAGG GAATTGTGGC
     50 1001 AGAAATCAAC ACCAAGACTG GTGAATCTGA ATGCCGCTAT TATAAAGAGC
     1051 GGCTTCTTTA CTTGGAAGAA GGCCAGAAAG ACTCCCTGAT CGACAGCTCT
     1101 CGAGTCTTGT GTTGTTCATGG TGAAGTGAAG AACAATCGGG GCGTGAAACT
     1151 GCATGTTTTT CTGTTCCAAG AAGTGCTTGT GATCACTCGA GCCGTCAACC
     1201 ACAATGAGCA GCTTTGCTAC CAGCTGTACC GTCAGCCAAT CCCCCTGAAA
     55 1251 GACCTCCTGC TGGAAGACCT CCAGGATGGA GAAGTGAGGC TGGGTGGCTC
     1301 CCTGCGAGGG GCATTGAGCA ACAATGAGAG AATTAAAAAC TTCTTCAGAG
     1351 TCAGTTTCAA AAATGGATCC CAAAGTCAGA CCCACTCGCT ACAAGCCAAT
     1401 GACACTTTCA ACAAACAGCA GTGGCTTAAC TGTATTGCTC AAGCCAAAGA

```

1451 AACAGTTTTG TGTGCTGCCG GGCAAGCTGG GGTGCTTGAC TCCGAGGGAT
1501 CGTTCCTAAA TCCCACCACC GGGAGCAGAG AGCTACAGGG AGAAAACAAAA
1551 CTTGAGCAGA TGGACCAATC GGACAGTGAG TCAGACTGTA GTATGGACAC
1601 GAGTGAGGTC AGCCTCGACT GTGAGCGCAT GGAACAGACA GACTCTTCCT
5 1651 GTGGAAACAG CAGGCACGGT GAAAGTAACG TCTGACAGAA GCATGTGCAC
1701 TTCGGGAAGC AGGCCTGCAT CTTACCTGTA CAGTATTTGC ATTCCACAGA
1751 TGGAACGGTT TGGAGAAGCA CTTTTTCATA CTTTTGTGAA AGTATACATG
1801 TTGGCCCAGT CTCTCGTATC TGTACCTTTG TCCCTAGTAC TGTAACCTGCC
1851 AATCTGTCTG TGTAAGCTGG AATCTGTGGC AACTATTACC CTGTGTTGTA
10 1901 TTTCCCAAGT GTCTGGATGG ATGGAGAGGT ACTCAAACAA GTTACTTTCA
1951 GTTGTCTCTG TGGATTTTAA AAAAATAGAA AAAGAATCTC AAAACTACTG
2001 TTTTACATAG ATTGTTTGAA GAGTCCTTCC TCTTGTGCTT CTGTACCACT
2051 TTCCCAGCTC TTAGATGTGG TAGCTAAAGG CACGGAATTT AGACGGCCTT
2101 GTAAATAGGG CATGAGGAAC TCATCTGTGT ATTGGGATGG TATTAGAGAG
15 2151 AGAATCAGGA AAGACCAACT CATGAAGTGA ACTTGGTTTG ATCTTACTCA
2201 ACTAGAAAGC TTGAAAACAT CCCTGGGGAT TCTGAAGGCT TAATTTTGCA
2251 AAGGAGGATG CATTGTCTGA ACTTTGCAAC TTCATCCAGT GCAAGTTTGA
2301 TGCAAGAATG TATTAGGACA TAAAATAGAG GCTGACCTTA AAAGGGCCAG
2351 GACAGAAGCG GCTGCCAGCT CTGAATCTTT AACTGAAATG CACATGGCAC
20 2401 CAGGAGGTGT CTCTCATAGT TGGTTGCTAG CCTAAAACAT CAGAATAGAA
2451 CCCAAAGGGC TTAGGAAGGC CTGCCAGGAT AACAAGAAGG CCCTGTATTC
2501 ATTGTGTTTC ATCTGCCTAG GCCTACTCAT TATTTTAGAG AATGAATGAA
2551 GCAACAAGGA AGAGAGACCA TGACTCTATC GATGACACTG TTTATAGAAA
2601 CACAGGAGAG GAAGAATTTG GAATGAAAAG CACTTCGTCA GAACCTTCTG
25 2651 TGGGAGCCAT TGAGAGAAAA GCATGGTCCA GTGCCTTCTG AGAAAGGCCA
2701 GAGCTTTGGG CTTTCCTGCT CTGCTTTTGG GTCGTCAATT TGCCATCTCT
2751 GGTTCTGTGC TATAATCAGA ATTGTAATTA TGTTCCTCCAG AGGCCAATTT
2801 CATTAACTCT GATTAATTAG AATCAGCTAG CCAGATTAGT AACCTCTTTG
2851 TCCAGCCTTG ATTTACAGTG CAGGGTAAAG TGCAGACCTT AAAACAGCT
30 2901 AAGTACCTAG AAGAGCTCCC TGCAAGTGTA AATATTAAGG ATGACCTGTG
2951 CAAAATTATA CCCACACCAG CACTAGTGGT AATTATTCTA AATTATTGCC
3001 AAAAAGTTTT TTTTAATCTG TCTTTCAAGT TTACAGAAAA GAAAGCAGTA
3051 AATGCATTGA TGTCATTTTA TTATGTACAT ATATCATGTG CATTCAAGCT
3101 GTGTGACAAG ATATATCAAT ATAAAAACAA GGTATATACT TTATTATTTT
35 3151 TTGAAAACAA GGATATTGTG ATCAATTTTA CCCTGTAAAA CATATTTCTG
3201 TATTTATAGG TCTTAAACAT GATGAATTTT TTCTATTACA AGTTTATTTA
3251 AAAGTGTCTT CTCAAGTCGT TATTGATACA GCAAGTGAAC CTGCTGCAGA
3301 CAGAAGCAGA GGAAAGCCAA GAACAGCCTT TATTGGTGAA GAAAAGAATG
3351 AATGATTCTT TGTAGGCGCC ATCAGCCACT TTTAGAAGCC ATCAGCCAGT
40 3401 GTGTTGGGAA AAGAGGTTTG TCAAGTGTG GCCTATGGGA AGGTGGTCAA
3451 TGAATGTTTT GATGAAATGA ATGTTTTTGT ATAATGGCCT TAACTTTTC
3501 TGGAAGTATT TCAAATAAAT TACATTATTA AGTCAAAAAA AAAAAAAAAA
3551 AAAAAAAAAA

45

BLAST Results

No BLAST result

50

Medline entries

55

98336196:

Alberts AS, Treisman R.; Activation of RhoA and SAPK/JNK
signalling
pathways by the

RhoA-specific exchange factor mNET1. EMB0 J 1998 Jul
15;17(14):4075-85

5

Peptide information for frame 3

10 ORF from 105 bp to 1682 bp; peptide length: 526
Category: strong similarity to known protein
Classification: Cell signaling/communication

15 1 MVAKDYPFYL TVKRANCSLE LPPASGPAKD AEEPSNKRVK PLSRVTS LAN
51 LIPPVKATPL KRFSQTLQRS ISFRSESRPD ILAPRPWSRN AAPSS TKRRD
101 SKLWSETFDV CVNQMLTSKE IKRQEAIFEL SQGEEDLIED LKLAKKAYHD
151 PMLKLSIMTE QELNQIFGTL DSLIPLHEEL LSQLRDVRKP DGSTEHV GPI
201 LVGWLPCLSS YDSYCSNQVA AKALLDHKKQ DHRVQDFLQR CLESPFSRKL
251 DLWNFLDIPR SRLVKYPLLL REILRHTPND NPDQQHLEEA INIIQGIVAE
20 301 INTKTGESEC RYYKERLLYL EEGQKDSLID SSRVLCCHGE LKNNRGVKLH
351 VFLFQEV LVI TRAVTHNEQL CYQLYRQPIP VKDLLLEDLQ DGEVRLGGS L
401 RGAFSNNERI KNFFRV SFKN GSQSQTHSLQ ANDTFNKQW LNCIRQAKET
451 VLCAAGQAGV LDSEGSFLNP TTGSRELQGE TKLEQMDQSD SESDCSMDTS
501 EVSLDCERME QTDSSCGNSR HGESNV

25

BLASTP hits

30 No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_29f24, frame 3

No Alert BLASTP hits found

35

Pedant information for DKFZphtes3_29f24, frame 3

Report for DKFZphtes3_29f24.3

40

[LENGTH] 560
[MW] 63202.85
[pI] 6.04
45 [HOMOL] TREMBL:AF094520_1 gene: "Net1"; product: "NET1
homolog"; Mus musculus NET1 homolog (Net1) mRNA, complete cds.
1e-162
[FUNCAT] 09.01 biogenesis of cell wall [S. cerevisiae,
YLR371w] 3e-16
50 [FUNCAT] 03.07 pheromone response, mating-type determination,
sex-specific proteins [S. cerevisiae, YLR371w] 3e-16
[FUNCAT] 10.02.09 regulation of g-protein activity [S.
cerevisiae, YLR371w] 3e-16
[FUNCAT] 09.04 biogenesis of cytoskeleton [S. cerevisiae,
55 YLR371w] 3e-16
[FUNCAT] 03.04 budding, cell polarity and filament formation
[S. cerevisiae, YLR371w] 3e-16

30	SEQ PRD	PPPGIVELGPPFAWEFC SRLGSAVTSQRAGPAAAMVAKDYFPYLT VKKRANCSLELPPASG ccccceeeccccccccchhhhhhhhhhhhhccccccccccccccccceeecccccccccccccccc
	SEQ PRD	PAKDAEEPSNKRVKPLSRVTS LANLIPPVKATPLKRFSQTLQRSISFRSESRPDILAPRP ccccccccccccccccccccccccccccccccccccchhhhhcccccccccccccccccccc
35	SEQ PRD	WSRNAAPSSTKRRDSKLWSETFDVCVNQMLTSKEIKRQEAIFELSQGEEDLIEDLKLAKK ccccccccccchhhhhhhhhhhhhccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
40	SEQ PRD	AYHDPMLKLSIMTEQELNQIFGTLDSL IPLHEELLSQLRDVRKPDGSTEHVGPILVGWLP hhhchhhccccccccccccceeecc
	SEQ PRD	CLSSYDSYCSNQVAAKALLDHKKQDHRVQDFLQRCLESPFSRKLDLWNFLDIPRSRLVKY ccccceeecccchhhhhhhhhhhhhcchhhhhhhhhhhccccccccccccceeeccccccccchh
45	SEQ PRD	PLLLREILRHTPNNDNPDAQHLEEA INIIQGIVAEINTKTGESECRYYKERLLYLEEGQKD hhhhhhhhhhccccccchhhhhhhhhhhhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhcccc
	SEQ PRD	SLIDSSRVL CCHGELKNNRGVKLVFLFQEV LVITRAVTHNEQLCYQLYRQPIPVKDLLL hhhhhhhheeeccccccccccccccccceeehhhhhhhhhhhhchhhhhhhhhhhhhcccccccccc
50	SEQ PRD	EDLQDGEVRLGGSLRGAFSNNERIKNFFRVSFKNQSQSTHSLQANDTFNKQQLNLCIRQ ccccccccccccccccchhhhhhhhhhhhhheeeccccchhhhhhhhhccccchhhhhhhhhhhhh
	SEQ PRD	AKETVLC AAGQAGVLDSEGSFLNPTTGSRELQGETKLEQMDQSDSESDCSMDTSEVSLDC hhhhhhhhhhccccceeeccccccccccccccccchhhhhhhhhhhhhhhcccccccccccccccccc
55	SEQ PRD	ERMEQTDSSCGNSRHGESNV cccccccccccccccccccccccc

(No Prosite data available for DKFZphtes3_29f24.3)

5 (No Pfam data available for DKFZphtes3_29f24.3)

DKFZphtes3_30pb

5 group: testis derived

DKFZphtes3_30pb encodes a novel 461 amino acid protein without similarity to known proteins.

10 No informative BLAST results; No predictive prosite, pfam or SCOP motife.

The new protein can find application in studying the expression profile of testis-specific genes.

15

similarity to C.elegans F41H10.4

perhaps complete cds.

20

Sequenced by LMU

Locus: unknown

25 Insert length: 1944 bp

Poly A stretch at pos. 1911, no polyadenylation signal found

30 1 GGAACAGACC ACTGGGCTGG CAGCTGAGTT GCAGCAGCAG CAGGCTGAGT
51 ACGAGGACCT TATGGGACAG AAAGATGACC TCAACTCCCA GCTCCAGGAG
101 TCATTACGGG CCAATAGTCG ACTGCTGGAA CAACTTCAAG AAATAGGGCA
151 GGAGAAGGAG CAGTTGACCC AGGAATTACA GGAGGCTCGG AAGAGTGCGG
201 AGAAGCGGAA GGCCATGCTG GATGAGCTAG CAATGGAAAC GCTGCAAGAG
251 AAGTCCCAGC ACAAGGAAGA GCTGGGAGCA GTTCGTCTAC GGCATGAGAA
35 301 GGAGGTGCTG GGGGTGCGTG CCCGCTATGA GCGTGAGCTC CGAGAGCTGC
351 ATGAAGACAA GAAGCGTCAG GAGGAGGAGC TCCGTGGGCA GATCCGGGAG
401 GAGAAGGCCC GGACACGGGA GCTGGAGACT CTCCAGCAGA CAGTGGGAAGA
451 ACTTCAAGCT CAGGTACATT CCATGGATGG AGCCAAGGGC TGGTTTGAAC
501 GCGGCTTGAA GGAAGCCGAG GAATCCCTGC AGCAGCAGCA GCAGGAACAA
40 551 GAGGAAGCCC TCAAGCAGTG TCGGGAGCAG CACGCTGCCG AGCTGAAGGG
601 CAAGGAGGAG GAGCTACAGG ATGTACGGGA TCAGCTCGAG CAGGCCCAGG
651 AGGAGCGGGA CTGCCACCTG AAGACCATTA GCAGCCTGAA GCAGGAGGTG
701 AAGGACACAG TGGATGGGCA GAGGATCCTG GAGAAGAAGG GCAGTGCTGC
751 GCTCAAGGAC CTCAAGCGGC AGCTGCATTT GGAGCGGAAA CGGGCAGATA
45 801 AGCTGCAGGA GCGACTGCAG GACATCCTCA CTAACAGCAA GAGCCGCTCA
851 GGCCTTGAGG AGCTGGTTCT CTCAGAGATG AACTCACCAA GCCGGACCCA
901 GACAGGGGAC AGCAGTAGCA TCTCCTCCTT CAGCTACCGG GAGATCTTGC
951 GGGAAAAGGA GAGCTCGGCT GTTCCAGCCA GGTCTTATC CAGCAGCCCT
1001 CAAGCCCAGC CCCCTCGGCC AGCAGAGCTG TCAGATGAGG AAGTGGCTGA
50 1051 GCTCTTTCAG CGGCTGGCAG AGACACAGCA GGAGAAATGG ATGCTGGAGG
1101 AGAAGGTGAA GCACCTGGAA GTGAGCAGTG CTTCCATGGC AGAGGACCTC
1151 TGCCGGAAGA GCGCCATCAT TGAGACCTAC GTCATGGACA GCCGGATCGA
1201 TGTGTCTGTG GCAGCAGGCC ACACAGACCG CAGCGGGCTG GGCAGCGTCC
1251 TGAGAGACCT AGTGAAGCCA GGCAGCAGGA ACCTTCGGGA GATGAACAAG
55 1301 AAGCTGCAGA ACATGCTGGA GGAGCAGCTC ACCAAGAATA TGCACTTGCA
1351 CAAGGATATG GAAGTTCTGT CCCAGGAAAT TGTGCGGCTC AGCAAGGAGT
1401 GCGTGGGGCC TCCTGACCCA GACCTAGAGC CAGGAGAAAC CAGCTAAAGA
1451 CCTGCAGGCT GCACCCACCT CCTCCCCTTC CTACCCCTTA GGATGCTATT

1501 CCCTTGGGCT GTGGTGGAAA AATGAGGGCT GGAGCCAAAA TCAAATAGCT
1551 TGGGAGACTG GACATTAAAG GGGCTAGAGG CCTGATGGTT AGTGTTAATG
1601 ATCCTGTCTT AGGGCAGAGG CCACCAGGGA GTGGGGATCC TGAGGGAAGG
5 1651 GGCAGGGATT TCTCCTTCTT CTTGGTCCTG GCTCCCAAGG GCTTCTGTCT
1701 TCATCTCTGC ATGAGCTCTC CTTCCCAGAG ACCAACTCTT TTTTATTTTA
1751 TTTTATTTTT TAATTTATGT CTGGAGCCTG GCTACTCTGC ATTTGGGATT
1801 GGGGATGCTG GGTGGGTGTG TGGTCCATGT TCAGCGTTCT AGCAACACGT
1851 GTGTGTGTGT GTGTGTAAAG GCTATGCAGC CAAAATACCA TCTGGCCAGA
1901 CGGGCCCACC CACAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAG

10

BLAST Results

15 No BLAST result

Medline entries

20

No Medline entry

25

Peptide information for frame 2

ORF from 62 bp to 1444 bp; peptide length: 461

Category: similarity to unknown protein

30 Classification: no clue

35

1 MGQKDDLNSQ LQESLRANSR LLEQLQEIGQ EKEQLTQELQ EARKSAEKRK
51 AMLDELAMET LQEKSQHKEE LGAVLRHEK EVLGVRARYE RELRELHEDK
101 KRQEEELRGQ IREEKARTRE LETLQQTVEE LQAQVHSMDS AKGWFERRLK
151 EAEESLQQQQ QEQEEALKQC REQHAAELKG KEEELQDVDR QLEQAQEERD
201 CHLKTISSLK QEVKDTVQGG RILEKKGSAA LKDLKRQLHL ERKRADKLQE
251 RLQDILTNSK SRSGLEELVL SEMNSPSRTQ TGDSSSISSF SYREILREKE
301 SSAVPARSL SPPQAQPPRP AELSDEEVAE LFQRLAETQQ EKWMLEEKVK
351 HLEVSSASMA EDLCRKSII ETYVMSRID VSVAAGHTDR SGLGSVLRDL
401 VKPGDENLRE MNKKLQNMLE EQLTKNMHLH KQMEVLSQEI VRLSKECVGP
451 PDPDLEPGET S

40

45

BLASTP hits

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_30pb, frame 2

50

No Alert BLASTP hits found

Pedant information for DKFZphtes3_30pb, frame 2

55

Report for DKFZphtes3_30pb.2

-370-

.....CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC.....

5

10 COILS

15 (No Pfam data available for DKFZphtes3_30pb.2)

DKFZphtes3_31a10

5 group: nucleic acid management

DKFZphtes3_31a10 encodes a novel 542 amino acid protein with similarity to histone H1 of *Drosophila hydei*.

10 Histone H1 variants are known to act as specific regulators of genes via the differential condensation of DNA.

The new protein can find application in modulating/blocking the transcriptional activity and in expression profiling.

15

weak similarity to *Drosophila* histone H1

perhaps complete cds.

20

Sequenced by LMU

Locus: /map="13"

25 Insert length: 2887 bp

Poly A stretch at pos. 2855, polyadenylation signal at pos. 2839

30 1 AGATGATCCC CAAAGTCAAC ATATGACATT AAGCCAGGCA TTTCACCTTA
51 AAAACAATAG TAAAAAGAAA CAAATGACTA CAGAAAAACA AAAGCAAGAT
101 GCTAACATGC CCAAGAAACC TGTGCTTGGA TCTTATCGTG GCCAGATTGT
151 TCAGTCTAAG ATTAATTCAT TTAGAAAACC TCTACAAGTC AAAGATGAGA
201 GTTCTGCAGC AACAAAGAAA CTTTCAGCCA CTATACCTAA AGCCACAAAA
251 CCTCAGCCTG TAAACACCAG CAGTGTAACA GTGAAAAGTA ATAGATCCTC
35 301 CAATATGACT GCCACTACTA AATTTGTGAG CACTACATCT CAGAACACAC
351 AACTTGTGCG ACCTCCTATT AGAAGTCATC ACAGTAATAC CCGGGACACT
401 GTGAAACAAG GCATCAGTAG AACCTCTGCC AATGTTACAA TCCGGAAAGG
451 GCCTCATGAA AAAGAACTAT TACAATCAAA AACAGCTTTA TCTAGTGTCA
501 AAACCAGTTC TTCTCAAGGT ATAATAAGAA ATAAGACTCT ATCAAGATCC
40 551 ATAGCATCTG AAGTTGTAGC CAGGCCTGCT TCATTGTCTA ATGATAAACT
601 GATGGAAAAG TCAGAGCCCG TTGACCAGCG AAGACATACT GCAGGAAAAG
651 CAATTGTTGA TAGTAGATCA GCTCAGCCCA AAGAAACCTC GGAAGAGAGA
701 AAAGCTCGTC TGAGTGAGTG GAAAGCTGGC AAAGGAAGAG TGCTAAAAAG
751 GCCCCCTAAT TCAGTAGTTA CTCAGCATGA GCCTGCAGGA CAAAATGAAA
45 801 AACTAGTTGG GTCTTTTGG ACTACCATGG CAGAAGAAGA TGAACAAAGA
851 TTATTTACTG AAAAAAGTAA CAACACATTT TCTGAATGCC TGAACCTGAT
901 TAATGAGGGA TGTCCAAAAG AAGATATACT GGTCACTCTG AATGACCTGA
951 TTAATAATAT TCCAGATGCC AAAAAAGCTTG TTAAGTATTG GATATGTCTT
1001 GCACCTATTG AACCAATCAC AAGTCCTATT GAAAATATTA TTGCAATCTA
50 1051 TGAGAAAGCC ATTCTGGCAG GGGCTCAGCC TATTGAAGAG ATGCGACACA
1101 CGATTGTAGA TATTCTAACA ATGAAGAGTC AAGAAAAAGC TAATTTAGGA
1151 GAAAATATGG AGAAGTCTTG TGCAAGCAAG GAAGAAGTCA AAGAAGTCAG
1201 TATTGAAGAT ACAGGTGTTG ATGTAGATCC AGAAAACTG GAAATGGAGA
1251 GTAAACTTCA TAGAAATTTG CTATTTCAAG ATTGTGAAAA AGAGCAAGAC
55 1301 AACAAAACAA AAGATCCAAC CCATGATGTT AAAACCCCCA ATACAGAAAC
1351 GAGGACAAGT TGCTTAATTA AATATAATGT GTCTACTACG CCATACTTGC
1401 AAAGTGTGAA AAAAAAGGTG CAGTTTGATG GAACAAATTC CGCATTTAAA
1451 GAGCTGAAGT TTTTAACACC AGTGAGACGT TCTCGACGTC TTCAAGAGAA

1501 AACTTCTAAA TTGCCAGATA TGTAAAAAGA TCATTATCCT TGTGTGTCTT
 1551 CATTGGAACA GCTAACGGAG TTGGGAAGAG AAAGCTGATGC TTTTGTATGC
 1601 CGCCCTAATG CAGCACTGTG CCGGGTGTAC TATGAGGCTG ATACAACATA
 1651 AGAGAAATAA AGCTCTGTAA GGAATGGGG TTTTATTAT TGTGGGGTG
 5 1701 TTTTGTTTTG AGTAGCTTTA TATTGCTCTT AGGTCTGGAG TTGGCCATGT
 1751 ACCTATGTAT CCTAAGCATT CACGGCAGTG AGCTCCTTTA CTAACATTCA
 1801 TGTATGGCA AGAGTTGTCC TCTACATTGG AAAGCTAATC CTACCTTGTG
 1851 AGTTTCAACC AACTGAGTTT TTTCTTTAAG AAAGGTAAAT TTTGTGAGCT
 1901 AGTTTACTAT GTTCCTTGAA TATAAACAGG TTATAATACT ACCCTGTTCA
 10 1951 CTTTACTAAA TATAAGTACA GTAATGATGC ATAATTAGAA AATGAGGTAT
 2001 TCTAGGTAAG ATGTATGTTT GCCTTGACAT GTTTTAAAAA GTTATGATGT
 2051 ACCTCCCTGC CTTTAAACAG AATACTTTTT TCTTTTTTTT GGCCTTTCTC
 2101 AGATTAGTCA AAAATTCTAT AGAATGACTC ACTTCGAATA CTAAGACACA
 2151 GGAGGTTTAG CCGCTTTCT TACCAAATTC ATGTTACCCA GACTTGTGTT
 15 2201 CTCTTGCGTC CCTTGGACTG CCTGTTGATT GATGGAAAGT GTCTGCACTG
 2251 ACACTTTTCG TCAGTAGTCT GTAGTTTCGT GGCCTCTTTT GATTATAACT
 2301 GGGGTCACCA AGAAGGTTTA CTTAATTAAA TACCGCATT TTAAGAGAAG
 2351 ATACTTTGTG TAAGAAAAGA TGCCACATTT AGTGGTTTAA CTTTGTAAAC
 2401 TTCACCTGAT AGTTTTTAAAG CAATTAGAAT GGAGTAGGG AAAGAACATA
 20 2451 TCATACTGAA CAAATGTCAT TCTAGTTTAG ATAGCATTTC TAAGATAACT
 2501 GATACTAATA CTTGTTTTCT TCCCTATAAC ATAAAAAACT TCACTGTTAA
 2551 GTCATGTCCC TTGAAACATG ATAGTTACAT ACACAGTTTT CTCTCCACAC
 2601 ATAAATAACA CCACTAAAGT TGTTTTGTAA GGTTCCAAAC TAATATGGCA
 2651 TATATCAACT CTACAGTTTC AAATAAATGA CTTTTTAATT GTAAAAGATT
 25 2701 AGTTGAAAAA CTGTATGAAT GTGAAGATCA CATGCTTAGT CATTTTTATG
 2751 TTCATTCCAC TTTGTATATC TTTTCTATTT ATTGACTTCT CATGTTCTAG
 2801 AGAGTAGGAC TTTTATTCCG TGTACCTGAT ATATATACAA TTAATAATATC
 2851 TGTATAATTA AAAAAAAAAA AAAAAAAAAA AAAAAAG

30

BLAST Results

No BLAST result

35

Medline entries

40 No Medline entry

Peptide information for frame 2

45

ORF from 23 bp to 1648 bp; peptide length: 542
 Category: similarity to known protein
 Classification: unclassified

50

55

1 MTLSQAFHLK NNSKKKQMTT EKQKQDANMP KKPVLGSYRG QIVQSKINSF
 51 RKPLQVKDES SAATKKLSAT IPKATKPQPV NTSSVTVKS N RSSNMTATTK
 101 FVSTTSQNTQ LVRPPIRSHH SNTRDTVKQG ISRTSANVTI RKGPHEKELL
 151 QSKTALSSVK TSSSQGIIRN KTLRSRIASE VVARPASLSN DKLMEKSEPV
 201 DQRRHTAGKA IVDSSRAQPK ETSEERKARL SEWKAGKGRV LKRPPNSVVT
 251 QHEPAGQNEK LVGSFWTTMA EEDEQRLFTE KVNNTFSECL NLINEGCPKE
 301 DILVTLNLI KNIPDAKKLV YWICLALIE PITSPIENII AIYEKAILAG
 351 AQPIEEMRHT IVDILTMKSQ EKANLGENME KSCASKEEVK EVSIEDTGVD

401 VDPEKLEMES KLHRNLLFQD CEKEQDNKTK DPTH DVKTPN TETR TSCLIK
451 YNVSTTPYLQ SVKKKVQFDG TNSAFKELKF LTPVRRSRRL QEKTSKLPDM
501 LKDHYPVCVSS LEQLTELGRE TDAFVCRPNA ALCRVVYEAD TT

5

BLASTP hits

No BLASTP hits available

10

Alert BLASTP hits for DKFZphtes3_31a10, frame 2

No Alert BLASTP hits found

15

Pedant information for DKFZphtes3_31a10, frame 2

Report for DKFZphtes3_31a10.2

20

[LENGTH] 549
[MW] 61677.36
[pI] 9.33
[KW] Alpha_Beta
25 [KW] LOW_COMPLEXITY 2.19 %

30 SEQ DDPQSQHMTLSQAFHLKNNSSKKKQMTTEKQKQDANMPKKPVLGSYRGQIVQSKINSFRKP
SEGxxxxxxxxxxxx.....
PRD cccccchhhhhheeeccccccccchhhhhhhhhcccccccccccccccccccccccc

35 SEQ LQVKDESSAATKKLSATIPKATKPQPVNTSSVTVKSNRSSNMTATTKFVSTTSQNTQLVR
SEG
PRD cccccchhhhhhhhhhhcc

40 SEQ SRSIASEVVARPASLSNDKLMKSEPVDQRRHTAGKAIVDSRSAQPKETSEERKARLSEW
SEG
PRD hhhhhhheeeccccchhhhhhhccccchhhhhhhcceeccccccccccccchhhhhhhhhhh

45 SEQ KAGKGRVLKRPPNSVVTQHEPAGQNEKLVGSFWTTMAEEDQRLFTEKVNNTFSECLNLI
SEG
PRD hccccccccccccccccccccccccccccccccccccchhhhhhhhhhhhhhhhhcccccccc

50 SEQ NEGCPKEDILVTLNDLIKNIPIAKKLVKYWICLALIEPITSPIENIIAIYEKAILAGAQP
SEG
PRD cccccccccccccccccccccccccchhhhhhhhhhhhhccccccccchhhhhhhhhhhhhcchh

55 SEQ IEEMRHTIVDILTMKSQEKANLGENMEKSCASKEEVKEVSIEDTGVDVDPEKLEMESKLH
SEG
PRD hhhhhhhhhhhhhhhhhhhhhhhhhccchhhhhccccccccccccccccccccchhhhhhhhh

SEQ RNLLFQDCEKEQDNKTKDPTH DVKTPNTETR TSCLIKYNVSTTPYLQSVKKKVQFDGTNS
SEG
PRD cchhhhhhhheeeccccch

SEQ AFKELKFLTPVRRSRRLQEKTSKLPDMLKDHYPVSSLEQLTELGRETD AFVCRPNAALC
SEG
PRD hhhhhhchhhhhhhhhhhhhhhccccccccccccchhhhhhhhccccceeeccceee

5

SEQ RVYYEADTT
SEG
PRD eeeccccc

10

(No Prosite data available for DKFZphtes3_31a10.2)

(No Pfam data available for DKFZphtes3_31a10.2)

DKFZphtes3_31j20

5 group: signal transduction

DKFZphtes3_31j20 encodes a novel 392 amino acid protein that contains a Protein phosphatase 2C motif.

10 The novel protein shares 95% identity with the rat protein phosphatase 2C and is expressed ubiquitously. PP2C is a structurally diversified protein phosphatase family with a wide range of functions in cellular signal transduction. The transcription of the PP2Cdelta gene was activated in response to stress, like alcohol or UV irradiation. PP2C plays a role in cell cycle control.

20 The new protein can find application in and the diagnosis/therapy of stress related diseases and cancer, as well as a for modulation of cell cycle and signal transduction.

strong similarity to protein phosphatase 2C (*Rattus norvegicus*)

25 Sequenced by LMU

Locus: unknown

Insert length: 1436 bp

30 Poly A stretch at pos. 1367, polyadenylation signal at pos. 1341

```

      1 CGCTGCTCGC GGGCTGAGTG TCTGTCGCTG CTGCCGCCTC CACCCAGCCT
    51 CCGCCATGGA CCTCTTCGGG GACCTGCCGG AGCCCGAGCG CTCGCCGCGC
  35 101 CCGGCTGCCG GGAAGAAGAG TCAGAAAGGA CCCCTGCTCT TTGATGACCT
    151 CCCTCCGGCC AGCAGTACTG ACTCAGGATC AGGGGGACCT TTGCTTTTGT
    201 ATGATCTCCC ACCCGCTAGC AGTGCGGATT CAGGTTCTCT TGCCACATCA
    251 ATATCCAGAG TGGTAAAGAG TGAAGGGAAA GGAGCAAAGA GAAAAACCTC
    301 CGAGGAAGAG AAGAATGGCA GTGAAGAGCT TGTGGAAAAG AAAGTTTGTA
  40 351 AAGCCTCTTC GGTGATCTTT GGTCTGAAGG GCTATGTGGC TGAGCGGAAG
    401 GGTGAGAGGG AGGAGATGCA GGATGCCCAC GTCATCCTGA ACGACATCAC
    451 CGAGGAGTGT AGGCCCCCAT CGTCCCTCAT TACTCGGGTT TCATATTTTG
    501 CTGTTTTTGA TGGACATGGA GGAATTCGAG CCTCAAAATT TGCTGCACAG
    551 AATTTGCATC AAAACTTAAT CAGAAAATTT CTAAGGAG ATGTAATCAG
  45 601 TGTAGAGAAA ACCGTGAAGA GATGCCTTTT GGACACTTTC AAGCATACTG
    651 ATGAAGAGTT CCTTAAACAA GCTTCCAGCC AGAAGCCTGC CTGGAAAGAT
    701 GGGTCCACTG CCACGTGTGT TCTGGCTGTA GACAACATTC TTTATATTGC
    751 CAACCTCGGA GATAGTCGGG CAATCTTGTT TCGTTATAAT GAGGAGAGTC
    801 AAAAAATGTC AGCCTTAAGC CTCAGCAAAG AGCATAATCC AACTCAGTAT
  50 851 GAAGAGCGGA TGAGGATACA GAAGGCTGGA GGAAACGTCA GGGATGGGCG
    901 TGTTTTGGGC GTGCTAGAGG TGTACGCTC CATTGGGGAC GGGCAGTACA
    951 AGCGCTGCGG TGTACCTCT GTGCCGACA TCAGACGCTG CCAGCTGACC
  1001 CCCAATGACA GGTTCATTTT GTTGGCCTGT GATGGGCTCT TCAAGGTCTT
  1051 TACCCAGAA GAAGCCGTGA ACTTCATCTT GTCCTGTCTC GAGGATGAAA
  1101 AGATCCAGAC CCGGGAAGGG AAGTCCGAG CCGACGCCC GTCAGGAAGCA
  1151 GCCTGCAACA GGCTGGCCAA CAAGCGGTG CAGCGGGGCT CAGCCGACAA
  1201 CGTCACTGTG ATGGTGGTGC GGATAGGGCA CTGAGGGGTG GCGCGCGGCC
  1251 AGGAGCACGC ATGGTATTGA CTAAAAGGT TCATTTTGTG TGTGTGCACA

```

1301 TTGTGTGTTT TGTGTACTCC TGTGGGACTC CCATGGTTGT AAATAAAGGT
1351 TTCTCTTTT TTTTCCTAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
1401 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAG

5

BLAST Results

No BLAST result

10

Medline entries

15 99074314:

Tong Y, Quirion R, Shen SH.; Cloning and characterization of a
novel
mammalian PP2C
isozyme. J Biol Chem 1998 Dec 25;273(52):35282-90

20

Peptide information for frame 2

25

ORF from 56 bp to 1231 bp; peptide length: 392
Category: strong similarity to known protein
Classification: Protein management
Prosit motifs: PP2C (147-155)

30

1 MDLFGDLPEP ERSRPAAAGK EAQKGPLLFD DLPPASSTD S GSGGPLLFDD
51 LPPASSGDSG SLATSIQMV KTEGKGAKRK TSEEEKNGSE ELVEKKVCKA
35 101 SSVIFGLKGY VAERKGEREE MQDAHVILND ITEECRPPSS LITRVSYFAV
151 FDGHGGIRAS KFAAQNLHQ N LIRKFPKGDV ISVEKTVKRC LLDTFKHTDE
201 EFLKQASSQK PAWKDGSTAT CVLAVDNILY IANLGDSRAI LCRYNEESQK
251 HAALSLSKEH NPTQYEERM R IQKAGGNVRD GRVLGVLEVS RSIGDGQYKR
301 CGVTSVPDIR RCQLTPNDRF ILLACDGLFK VFTPEEAVNF ILSCLEDEKI
40 351 QTREGKSAAD ARYEAACNRL ANKAVQRGSA DNVTVMVVRI GH

BLASTP hits

45

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_31j20, frame 2

50 No Alert BLASTP hits found

Pedant information for DKFZphtes3_31j20, frame 2

55

Report for DKFZphtes3_31j20.2

[LENGTH] 410

[MW] 44759.85
 [PI] 7.95
 [HOMOL] TREMBL:AF095927_1 product: "protein phosphatase 2C"; Rattus norvegicus protein phosphatase 2C mRNA, complete cds.
 5 0.0
 [FUNCAT] 03.01 cell growth [S. cerevisiae, YDL006w] 6e-25
 [FUNCAT] 10.03.13 key phosphatases [S. cerevisiae, YDL006w] 6e-25
 10 [FUNCAT] 09.16 mitochondrial biogenesis [S. cerevisiae, YDL006w] 6e-25
 [FUNCAT] 11.01 stress response [S. cerevisiae, YDL006w] 6e-25
 [FUNCAT] 03.04 budding, cell polarity and filament formation [S. cerevisiae, YDL006w] 6e-25
 15 [FUNCAT] 01.05.04 regulation of carbohydrate utilization [S. cerevisiae, YDL006w] 6e-25
 [FUNCAT] 98 classification not yet clear-cut [S. cerevisiae, YER089c] 1e-23
 [FUNCAT] 99 unclassified proteins [S. cerevisiae, YOR090c] 1e-12
 20 [FUNCAT] 03.22 cell cycle control and mitosis [S. cerevisiae, YJL005w] 3e-10
 [FUNCAT] 03.10 sporulation and germination [S. cerevisiae, YJL005w] 3e-10
 [FUNCAT] 30.02 organization of plasma membrane [S. cerevisiae, YJL005w] 3e-10
 25 [FUNCAT] 01.03.10 metabolism of cyclic and unusual nucleotides [S. cerevisiae, YJL005w] 3e-10
 [FUNCAT] 10.04.03 second messenger formation [S. cerevisiae, YJL005w] 3e-10
 30 [BLOCKS] PRO1023F
 [BLOCKS] PRO0677D
 [BLOCKS] BL01032I
 [BLOCKS] BL01032H
 [BLOCKS] BL01032G
 35 [BLOCKS] BL01032C Protein phosphatase 2C proteins
 [BLOCKS] BL01032B Protein phosphatase 2C proteins
 [SCOP] dlabq_ 4.98.1.1.1 Protein serine/threonine phosphatase 2C [Huma] 1e-107
 [EC] 3.1.3.43 [Pyruvate dehydrogenase (lipoamide)]-phosphatase 3e-09
 40 [EC] 3.1.3.16 Phosphoprotein phosphatase 7e-35
 [EC] 4.6.1.1 Adenylate cyclase 2e-11
 [PIRKW] duplication 5e-11
 [PIRKW] tandem repeat 8e-09
 45 [PIRKW] serine/threonine-specific phosphatase 2e-27
 [PIRKW] magnesium 6e-26
 [PIRKW] cAMP biosynthesis 5e-11
 [PIRKW] liver 2e-27
 [PIRKW] leucine zipper 1e-08
 50 [PIRKW] mitochondrion 3e-09
 [PIRKW] phosphoric monoester hydrolase 7e-35
 [PIRKW] phosphorus-oxygen lyase 2e-11
 [SUPFAM] leucine-rich alpha-2-glycoprotein repeat homology 2e-11
 55 [SUPFAM] yeast adenylate cyclase catalytic domain homology 2e-11
 [SUPFAM] kinase interaction domain homology 3e-11
 [SUPFAM] yeast adenylate cyclase 5e-11

[illegible]

30 PSQ1032 165->174 PP2C PD0C00792

```

40 HMM *G1CcMQGPRWRMsMEDaHiaylNF.....pcnldWWhiMFFGVFDGHg
      +++ +G R++M+DAH+ + ++ P++L ++
      +++F+VFDGHG
      Query 128 YVAERKG--EREEMQDAHVILNDITEECRPPSSLITR-
45 VSYFAVFDGHG 173

      HMM GDQCSQWCgeHWHdII*
      G+++S++ +++H+ +
      Query 174 GIRASKFAAQNLHQNL 189
50

```

DKFZphtes3_5k22

5. group: signal transduction

DKFZphtes3_5k22 encodes a novel 455 amino acid protein with similarity to human paraneoplastic neuronal antigen MA1.

10 Antibodies against MA1 were found in patients with paraneoplastic neurological disorders. The protein is predominantly expressed in testis and brain, but ESTs are also found in liver, lung, uterus and kidney.

15 The new protein can find application in studying/therapy of paraneoplastic neurological disorders.

strong similarity to paraneoplastic neuronal antigen MA1

20 Sequenced by Qiagen

Locus: unknown

25 Insert length: 3534 bp

Poly A stretch at pos. 3514, polyadenylation signal at pos. 3494

```

30      1 GAACGTCCGC GCTGGGAGCC AGGGGTGCCC GACCCCCGTC CGCCGCCGCC
      51 GCCGCCGCCG CGCATAGCCC CCGGAGAGCC CTCTGGGGAC CCCGACCAGA
      101 AGGGACCTTG CCCTGGGAGA AGGCTGTGGA GACCTGGGCC TTCTGCGATC
      151 ACCCTAGGAG TTGATCCAGA TATGTGCCTC ACGCCCTGAT CACTCCCCCC
      201 AAATTAGTAT CCGCAGAGAT TCGAGGACAT GCCGTTGACC TTGTTACAGG
      251 ACTGGTGTCT GGGGGAACAC CTGAACACCC GGAGGTGCAT GCTCATCCTG
35      301 GGGATCCCCG AGGACTGTGG CGAGGATGAG TTTGAGGAGA CACTCCAGGA
      351 GGCTTGCCAG CACCTGGGCA GATACAGGGT GATTGGCAGG ATGTTTAGGA
      401 GGGAGGAGAA CGCCAGGGCG ATTCTACTGG AGCTGGCACA AGATATCGAC
      451 TATGCTTTGC TCCCAAGGGA AATACCAGGA AAGGGGGGGC CCTGGGAAGT
      501 GATTGTAAAA CCCCCTAACT CAGATGGGGA ATTTCTCAAC AGACTGAACC
40      551 GCTTCTTAGA GGAGGAGAGG CGGACCGTGT CAGATATGAA CCGAGTCCTC
      601 GGGTCGGACA CCAATTGTTC GGCTCCAAGA GTGACTATAT CACCAGAGTT
      651 CTGGACCTGG GCCCAGACTC TGGGGGCGAG AGTGCAGCCT CTGCTAGAAC
      701 AAATGTTGTA CCGAGAACTA AGAGTGTTTT CTGGGAACAC CATATCCATC
      751 CCAGGTGCAC TGGCCTTTGA TGCCTGGCTT GAGCACACCA CTGAGATGCT
45      801 ACAGATGTGG CAGGTGCCCC AGGGGGAAAA GAGGCGGAGG CTGATGGAAT
      851 GCTTACGGGG CCCTGCTCTC CAGGTGGTCA GTGGGCTCCG GGCCAGCAAT
      901 GCTTCCATAA CTGTGGAGGA GTGCCTGGCT GCCTTGCCAG AGGTGTTCCG
      951 ACCTGTGGAG AGCCATAAAA TTGCCAGGT GAAGTTGTGT AAAGCCTATC
      1001 AGGAGGCAGG AGAGAAAGTA TCTAGCTTTG TGTTACGTTT GGAACCCCTG
50      1051 CTCCAAAGAG CTGTAGAAAA CAATGTGGTA TCACGTAGAA ACGTGAATCA
      1101 GACTCGCCTG AAACGAGTCT TAAGTGGGGC CACCCTTCCT GACAAACTCC
      1151 GAGATAAGCT TAAGCTGATG AAACAGCGAA GGAAGCCTCC TGGTTTCCTG
      1201 GCCCTGGTGA AGCTCCTGCG TGAGGAGGAG GAATGGGAGG CCACTTTAGG
      1251 TCCAGATAGG GAGAGTTGGG AGGGGCTGGA AGTAGCCCCA AGGCCACCTG
55      1301 CCAGGATCAC TGGGGTTGGG GCAGTACCCT TCCCTGCCCT TGGCAACAGT
      1351 TTTGATCGCA GGCCTTCCCA GGGCTACCGG CGCCGGAGGG GCAGAGGCCA
      1401 ACACCGAAGG GGTGGTGTGG CAAGGGCTGG CTCTCGAGGC TCAAGAAAC
      1451 GGAAACGCCA CACATTCTGC TATAGCTGTG GGAAGACGG CCACATCAGG
```

```

1501 GTACAGTGCA TCAACCCCTC CAACCTGCTC TTGGCCAAGG AGACAAAAGA
1551 GATATTGGAA GGAGGGGAAA GAGAAGCCCA GACAAACAGC AGATGAGTTG
1601 AGTGGGGCAG AGGGACAGGG CAGCCAGACC AAGGCCAAGC CTTCTCACCC
1651 TTGGCCAGCT GGAAGGGACT TCAGCAACCA AGACCACCTG GCAACAGGCT
5 1701 CAGTGGGGGT CAGGTCCAGG TCCCCGAAGA GGTGCTGGAG AGGAAAGCAG
1751 GGAGCCACTG CATCCAGCAC ATGGGGTGCC TGGGCCTCAG ATGGGGACCC
1801 CAAAGAAGCA GAAGCTGAAG AAGGTACGGC TGGGGGTCTT GTCTGTCTCA
1851 TCCAACCACC CCTAAATACC CACCCTGTGG ACTTTGAGCT GAACATGCCC
1901 ACTGGCCCCC AGGCCACATG GGACCTGGAG GAGCCTACCT GGGGCCTGCC
10 1951 CCTGCCAGCA GGTGCCAGGG CTGGTGAGGA AGAGCTGGGG GGCAGAGGTA
2001 AAGCCCTGCA GGGGAGGCCA CAGGGTCCAT CCCGTCTTCA GGATCATCTA
2051 CACTGCACTA GGGGAGCCCC AGGAAGGCAG CACCCTGGAG GCCCTGTGCC
2101 AGTGAGGACA GGAGACCCTA AGGCCCCGGG AGCCCAAGTC CAGCCAGAGG
2151 TTGTGCAGGC AAGGAGACCA AAGATTGATG AGAAGACCCC CAGCAGGGGT
15 2201 ACTGGGTACC CGGCAGGCCA GTGCCCTCAC AGTTGACTTG GACCAGGGTG
2251 GCTGTGAAGG GAAGTCTTTG TTGCAAAGGA GGAGGAGGAA AAGGGAGGAC
2301 TTGGTAGGGT TTTGTTTCTT CTGCTTGTTT CTGTACAGGG CCACCAGACT
2351 CCTGGAGAGA TCAAGCAAGG AGAACCTGGG GCTGCCATGG CCAAAGCAAC
2401 TCAACAGATG CCAATGCCAA TTCCAAGGCC AGCCACAACC CTGCCACCTT
20 2451 GGGGAATCCA GCCTGGAGGC ATCCCCTAAG CAGCCAGCCA TGGCCTGGGT
2501 GGAGGCACCT GAAGACGTCT GTCCCCAACT CCCCCAGCCC TGAGCTGGGA
2551 GATGACAGGG GGAAAGAGGC CCTCTCAAGG GTGCCAGATG CTTGGGTCTC
2601 CCAAGAGGGG TCCCCCAACT CACCCTTCCC GGGACAGGCT GCCCCTGTT
2651 CCAGGAAGCT CATCCTCAC TGTGTAGGCC CCTGTAGTGA CCCACGCGTC
25 2701 CAGCAGACGC CCACCCACCG CTAGCCGTTG TTCCTGTGCA AAGTAGTGTG
2751 CTATGCACCC ACCCAGGTGG CCGCCTCTGG GCCCAAGGCA CATGCTGTGA
2801 GCTTCCTGTG AGCCCAGGCT CTGCTCACTG CTGTCCCGCG TCATGAGCAC
2851 CACCTCTGCT TTCCCTGTGT AGATCTAGGC CAGTGGCTGC TTGTTCTTGT
2901 GGAGCTGTGT GTGTTCTTCT CTGAGCAGCT CCTCCCCGGA GTCCCCCAGC
30 2951 ACAGTCCAG GAGATGACAG GAAGGAAGCA CCAGGGCAAG GCGGACGCTC
3001 ACCCTGTGAC CACGATGGTG ACCGTGGCTG TGGGAGGAAG AACTGGACCC
3051 AGGACGGAGC GGGGCTGCCC TGCCTGAGGC TCCCGAGGAG CTTTGTGCTT
3101 TGGTGTTCCTA CCCCTGTTGT TACTCATGAC TCAGTTTCCT TGACCTGGTA
3151 GGGTGTTCCTA TGCTGTGTTT TCCAGTGTCC TGTGACTGTC CTGTGCGGGC
35 3201 CATAGGGCAG GGCCCTGCCC CAGCAGATGG GCTTGGGAGG GGGCTCCCTA
3251 AAGCCAGTGG ACACTGCCAG AGTCTACCTT CCTGGCAAGA GGCAGACCCC
3301 GGGGCCCTCA GGAAGGAGGG AGTTGGCAGC GGGGGCTGCA GCAGGAGTAG
3351 GAGCAGATGA GCGCTCTTGC CAGGAACCTC AGGAGGAGGG GGGCCGGGAC
3401 CTGTGTGGGA CCTGTGTCTT GTGGTGGCCG TTTGCAGTTT CTCTCTGTGT
40 3451 TGTGATTCCC TTCTCTTCAA TGGTTTCAGT ACGTGTCTTCT CTTCAATAAA
3501 CTTCAATTCAG TGTAAAAAAA AAAAAAAAAA AAAA

```

BLAST Results

No BLAST result

Medline entries

99158179:

Mal, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders.

Peptide information for frame 1

5

ORF from 229 bp to 1593 bp; peptide length: 455
 Category: strong similarity to known protein
 Classification: unclassified

```

10      1 MPLTLLQDWCR RGEHLNTRRC MLILGIPEDC GEDEFEEETLQ EACRHLGRYR
      51 VIGRMFRREE NAQAILLELA QDIDYALLPR EIPGKGGPWE VIVKPRNSDG
      101 EFLNRLNRFL EEERRTVSDM NRVLGSDTNC SAPRVTISPE FWTWAQTLGA
      151 AVQPLLEQML YRELRVFSGN TISIPGALAF DAWLEHTTEM LQMWQVPEGE
      201 KRRRLMECLR GPALQVVSG L RASNATITVE ECLAALQQVF GPVESHKIAQ
15      251 VKLCKAYQEA GEKVSSFVLR LEPLLQRAVE NNVSRRNVN QTRLKRVLSG
      301 ATLPDKLRDK LKLMKQRRKP PGFLALVKLL REEEWEATL GPDRESLEGL
      351 EVAPRPPARI TGVGAVPLPA SGNSFDARPS QGYRRRRGRG QHRRGGVARA
      401 GSRGSRKRKR HTFCYSCGED GHIRVQCINP SNLLAKETK EILEGGGEREA
      451 QTNSR

```

20

BLASTP hits

25 No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_5k22, frame 1

```

30 TREMBLNEW:AB020690_1 gene: "KIAA0883"; product: "KIAA0883
protein";
Homo sapiens mRNA for KIAA0883 protein, complete cds., N = 1,
Score =
722, P = 2.4e-71
35 TREMBL:AF037364_1 gene: "MAL"; product: "paraneoplastic neuronal
antigen MAL"; Homo sapiens paraneoplastic neuronal antigen MAL
(MAL)
mRNA, complete cds., N = 1, Score = 665, P = 2.6e-65

```

40

```

>TREMBLNEW:AB020690_1 gene: "KIAA0883"; product: "KIAA0883
protein"; Homo
sapiens mRNA for KIAA0883 protein, complete cds.
Length = 364

```

45

HSPs:

Score = 722 (108.3 bits), Expect = 2.4e-71, P = 2.4e-71
 Identities = 156/348 (44%), Positives = 215/348 (61%)

50

```

Query:      1
MPLTLLQDWCRGEHLNTRRCMLILGIPEDCGEDEFEEETLQEACRHLGRYRVIGRMFRREE 60
              M L LL+DWCR  ++ ++ +++ GIP D  E E +E LQE  +
LGRYR++G++FR++E
55 Sbjct:      1
MALALLEDWCRIMSVDEQKSLMVTGIPADFEEAEIQEVLQETLKSLGRYRLLGKIFRKQE 60

```

Query: 61
 NAQAILLELAQDIDYALLPREIPGKGGPWEVIVKPRNSDGXXXXXXXXXXXXXXXXXTVSDM 120
 NA A+LLEL +D D + +P E+ GKGG W+VI K N D
 TVS M

5 Sbjct: 61
 NANAVLLELLEDTDVSAIPSEVQGKGGVWKVIFKTPNQDTEFLERLNLFLKEKEGQTVSGM 120

Query: 121 NRVLGSDTNCAPRVTSPEFWTW--
 AQTGAAVQPLLEQMLYRELRFVSGNTISIPGAL 178
 R LG + A ISPE Q + A QPLL M YR+LRVFSG+

10 + P
 Sbjct: 121 FRALGQEGVSPATVPCISPELLAHLGQAMAHAPQPLLP-
 MRYRKLRVFSGSAVPAPEEE 179

15 Query: 179
 AFDAWLEHTTEMLQMWQVPEGEKRRRLMECLRGPALQVVSGLRASNASITVEECLAALQ 238
 +F+ WLE TE+++ W V E EK+R L E LRGPAL ++ ++A N
 SI+VEECL A +Q

20 Sbjct: 180
 SFEVWLEQATEIVKEWPVTEAEKKRWLAESLRGPALDLMHIVQADNPSISVEECLEAFKQ 239

Query: 239
 VFGPVESHKIAQVKLCKAYQEAAGEKVSSFVLRLEPLLQXXXXXXXXXXXXXXXXXXLKRVL 298
 VFG +ES + AQV+ K YQE GEKVS++VLRLE LL+

25 L++V+
 Sbjct: 240
 VFGSLESRRTAQVRYLKTYQEEGEKVSAYVLRLETLLRRRAVEKRAIPRRIADQVRLEQVM 299

Query: 299 SGATLPDKLRDKLKLKMKQRRKPPGFLALVKLLREEEEWEATLGPDRSLE
 348
 +GATL L +L+ +K + PP FL L+K++REEEE EA+ + ES+E

30 Sbjct: 300 AGATLNQMLWCRLRELKDQGPPPSFLELMKVIREEEEEEEASF--ENESIE
 347

35 Pedant information for DKFZphtes3_5k22, frame 1

Report for DKFZphtes3_5k22.1

40

45 [LENGTH] 455
 [MW] 51514.34
 [pI] 9.27
 [HOMOL] TREMBLNEW:AB020690_1 gene: "KIAA0883"; product:
 "KIAA0883 protein"; Homo sapiens mRNA for KIAA0883 protein,
 complete cds. 3e-75
 [BLOCKS] BL008768 Indoleamine 2,3-dioxygenase proteins
 [PFAM] Zinc finger, CCHC class

50 [KW] Alpha_Beta
 [KW] LOW_COMPLEXITY 13.41 %

55 SEQ MPLTLLQDWCRGEHLNTRRCMLILGIPEDCGEDEFEEETLQEACRHLGRYRVIGRMFRREE
 SEG
 PRD ccchhhhhccccccccccccccccccccccccchhhhhhhhhhhhhhhccceeehhhhhhhhhh

SEQ NAQAILLELAQDIDYALLPREIPGKGGPWEVIVKPRNSDGEFLNRLNRFLEEERRTVSDM

(No Prosite data available for DKFZphtes3_5k22.1)

```
35 HMM_NAME.. Zinc finger, CCHC-class ..
```

40 Query 412 TFCYSCGEDGHIRVQCIN 429

5 group: transmembrane protein

DKFZphtes3_7n12 encodes a novel 703 amino acid protein without similarity to known proteins.

10 The novel protein contains 1 transmembrane domain
No informative BLAST results; No predictive prosite, pfam or SCOP motif.

15 The new protein can find application in studying the expression profile of testis-specific genes and as a new marker for testicular cells.

putative protein

20 contains transmembrane domain
perhaps complete cds.

Sequenced by BMFZ

25 Locus: unknown

Insert length: 2347 bp

Poly A stretch at pos. 2271, polyadenylation signal at pos. 2253

30

```

      1 CGGCTGCAGT CTGGGCCGGG GCCCTGTGCC GCTGAAGACA TGGAGTTTGT
    51 GTCTGGATAC CGGGATGAGT TCCTTGATTT CACTGCCCTT CTCTTCGGCT
   101 GGTTCCGAAA GTTTGTGGCA GAGCGTGGAG CTGTAGGGAC TAGCCTTGAG
   35 151 GGCCGCTGCC GGCAGCTGGA GGCCAGATC AGAAGGCTAC CCCAGGACCC
      201 TGCCCTTTGG GTGCTCCATG TCCTGCCCAA CCATAGTGTG GGCATCAGCC
      251 TGGGGCAAGG GGCAGAACCA GGTCTTGGAC CAGGCCTGGG GACTGCCTGG
      301 CTCCTGGGAG ACAACCCTCC ACTCCACCTG CGAGACCTGA GCCCCTACAT
      351 CAGCTTTGTC AGCCTAGAGG ATGGGGAGGA AGGGGAGGAG GAAGAGGAGG
   40 401 AAGATGAAGA AGAAGAGAAG AGAGAGGACG GGGGTGCAGG CAGCACAGAG
      451 AAGGTGGAAC CAGAGGAGGA CCGGGAGCTA GCCCCTACCA GCAGGGAGTC
      501 CCCCCAGGAA ACAAAACCCTC CAGGAGAGTC AGAGGAGGCT GCCCAGGAGG
      551 CAGGAGGTGG CAAGGATGGC TGCCGAGAGG ACAGGGTGGA GAACGAAACA
      601 AGACCCCAAG AGAGGAAGGG ACAGAGGAGT GAGGCTGCCC CCTGACAGT
   45 651 TTCCTGTCTC TTACTTGTGA CGGATGAGCA TGGCACCATC TTGGGCATTG
      701 ATCTGCTAGT GGATGGAGCC CAGGGAACCG CAAGCTGGGG CTCAGGGACC
      751 AAGGACCTGG CTCCTTGGGC CTATGCTCTC CTCTGTCACA GCATGGCCTG
      801 TCCCATGGGC TCTGGGGATC CCCGAAAGCC CCGACAGCTT ACTGTGGGAG
      851 ATGCCCAGCT GCATCGAGAG CTGGAGAGCT TGGTCCCAAG GCTAGGTGTG
   50 901 AAGTTAGCCA AAACCCCAAT GCGGACATGG GGTCCCCGGC CAGGCTTCAC
      951 CTTTGCTTCC CTTCGTGCTC GAACCTGCCA TGTGTGTAC AGGCACAGCT
   1001 TTGAAGCGAA GCTGACACCT TGCCCCAGT GTAGTGCTGT CTTGTATTGT
   1051 GGAGAGGCTT GTCTCCGGGC TGACTGGCAG CGGTGCCAG ATGATGTGAG
   1101 TCACCGATTT TGGTGCCAA GGCTTGACAG CTTATGGAG CGGGACAGAG
   55 1151 AACTGGCAAC CCTACCTTTT ACCTACACCG CAGAGGTGAC CAGTGAACC
   1201 TTCAACAAAG AGGCCTTCCT GGCCTCTCGG GGCCTCACTC GTGGCTATTG
   1251 GACCCAGCTC AGCATGCTGA TTCCAGGCCC GGGCTTCTCC AGACACCCCC
   1301 GAGGCAACAC GCCATCCCTC AGCCTTCTTC GCGGTGGAGA CCCCTACCA
```

```

1351 CTTCTCCAGG GAGACGGGAC TGCCCTGATG CCTCCTGTGC CCCCACATCC
1401 ACCCCGGGGT GTTTTGTCC CTGAGCTCAA CATCCAAAAC AAACAGTCAC
1451 TGAAGATCCA CGTGGTGGAG GCCGGGAAGG AGTTTGACCT TGTATGGTG
5 1501 TTTTGGGAGC TTTTGGTCTT GCTCCCCCAT GTGGCCCTGG AGCTGCAGTT
1551 TGTAGGTGAT GGCCTGCCCC CCGAAAGCGA CGAGCAGCAT TTTACCCTGC
1601 AGAGGGACAG CCTGGAGGTG TCTGTCCGGC CTGGTTCCGG CATATCAGCA
1651 CGGCCAGCT CTGGCACTAA GGAGAAAGGG GGCCGCAGGG ACCTGCAGAT
1701 CAAGGTGTCA GCAAGGCCCT ACCACCTGTT CCAGGGGCCC AAGCCTGACC
1751 TGGTTATTGG ATTTAACTCC GGGTTTGCTC TCAAGGATAC GTGGCTGAGG
10 1801 TCTCTGCCCC GGTACAGTC CCTCCGAGTG CCAGCCTTCT TCACCGAGAG
1851 CAGCGAGTAC AGCTGTGTGA TGGACGGCCA GACCATGGCG GTGGCCACTG
1901 GAGGGGGCAC CAGCCCTCCC CAGCCCAACC CCTTCCGCTC CCCCTTTCGC
1951 CTCAGAGCGG CCGACAACCT CATGTCCTGG TACTGCAATG CCTTCATCTT
2001 CCACCTGGTT TACAAGCCTG CTCAGGGAG CGGGGCCCGC CCGGCGCCCG
15 2051 GGCCCCCACC CCCATCCCCA ACTCCCTCTG CTCCTCCTGC CCCCACCCGA
2101 AGGCGCCGAG GAGAAAAGAA ACCTGGGCGG GGGGCCCGCC GCGGGAATG
2151 AATGCTGATA CCCTAGTAGT CCCAGCTCC CAAACACTGA AAGGAAAACG
2201 TGAAAACACT CAAGGCCTAG GGGGAGGACA GGTGGTAAA ACATGAAAAG
2251 GTAAATAAAA TTAATTGTTT GAAAAAAAAA AAAAAAAAAA AAAAAAAAAA
20 2301 AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAA

```

BLAST Results

25

No BLAST result

Medline entries

30

No Medline entry

35

Peptide information for frame 1

```

ORF from 40 bp to 2148 bp; peptide length: 703
40 Category: putative protein
Classification: Transmembrane proteins unclassified

```

```

1 MEFVSGYRDE FLDFALLFG WFRKFVAERG AVGTSLEGRC RQLEAQIRRL
51 PQDPALWVLH VLPNHSVGIS LGQGAEPGPG PGLGTAWLLG DNPPLHLRDL
45 101 SPYISFVSLE DGEEGEEEE EDEEEKRED GGAGSTEKVE PEEDRELAPT
151 SRESPQETNP PGESEEAARE AGGGKDGCRE DRVENETRPQ KRKGQRSEAA
201 PLHVSCLLLV TDEHGTLGI DLLVDGAQGT ASWGSGBKDL APWAYALLCH
251 SMACPMGSGD PRKPRQLTVG DARLHRELES LVPRLGVKLA KTPMRTWGPR
301 PGFTFASLRA RTCHVCHRHS FEAKLTPCPQ CSAVLYCGEA CLRADWQRCP
50 351 DDVSHRFWCP RLAAFMERAG ELATLPFTYT AEVTSETFNK EAFLASRGLT
401 RGYWTQLSML IPGPGFSRHP RGNTPSLSLL RGGDPYQLLQ GDGTALMPPV
451 PPHPPRGV FV PELNIQNKQS LKIHVVEAGK EFDLVMVFWL LLVLLPHVAL
501 ELQFVGDGLP PESDEQHFTL QRDSELEVSVR PGSGISARPS SGTKEKGGRR
55 551 DLQIKVSARP YHLFQGP KPD LVIGFNSGFA LKDTWLRSLP RLQSLRVP AF
601 FTESSEYSCV MDGQTM AVAT GGGTSPPQPN PFRSPFRLRA ADNCMSWYCN
651 AFIFHLVYKP AQGSGARPAP GPPPPSPTPS APPAPTRRRR GEKKPGRGAR
701 RRK

```

5 No BLASTP hits available

No Alert BLASTP hits found

Report for DKFZphtes3_7n12.1

```

[LENGTH] 703
[MMW] 77312.72
[PI] 6.45
[KW] TRANSMEMBRANE 1
[KW] LOW COMPLEXITY 15.22 %

```

```
SEQ      MEFVSGYRDEFLDFTALLFGWFRKFVAERGAVGTSLEGRCRQLEAQIRRLPQDPALWVLH
SEG      .....
PRD      cccceecchhhhhhhhhhhhhhhhhhhhhcccccccchhhhhhhhhhhhhhhcccccccccc
MEM      .....
```

```
SEQ      VLPNHSVGISLGQGAEPGPGPLGTAWLLGDNPPLHLRDLSPYISFVSLEDGEEGEEEEEE  
SEG      .....xxxxxxxxxxxxxx  
PRD      cccccccccccccccccccceeeeeeccccccccccccccccceeeeecccchhhhhhhh  
MEM      .....
```

[illegible]

```
SEQ  DRVENETRPQKRKGQRSEAAPLHVSCLLLVTDEHGTILGIDLLVDGAQGTASWGSGTKDL
SEG  .....
PRD  eeeeeeeeeeeeeeeeeeechhhhhhheeeccccccccchhhhhhcccccccccccccccc
MEM  .....
```

```
SEQ      APWAYALLCHSMACPMGSGDPRKPRQLTVGDARLHRELESLVPRLGVKLAKTPMRTWGR
SEG      .....
PRD      hhhhhhhhhhhhhhhcccccccccccccccccccccccccccccccccccccccccccc
MEM      .....
```

```
SEQ PGFTFASLRARTCHVCHRHSFEAKLTCPQCSAVLYCGEACLRADWQRCPPDDVSHRFWCP
SEG .....
PRD cccccchhhhhhhhhccccccccccccccccccccccccccccccccccccccccch
MEM .....
```

```
SEQ  RLAAFMERAGELATLPFTYTA EVTSETFNKEAFLASRGLTRGYWTQLSMLIPGPGFSRHP
SEG  .....
PRD  hhhhhhhhhhhhhhhccccccccchhhhhhhhhhhhhhhhhccccchhhhhcccccccccccc
MEM  .....
```

```

SEQ  RGNTPSLSLLRGGDPYQLLQGDGTALMPPVPPHPPRGVFPVPELNIQNKQSLKIHVVEAGK
SEG  .....xxxxxxxxxxxxxxxx.....
PRD  cccccceeeeeeccccceccccccccccccccccccccccccchhhhhheeeeeeccc
MEM  .....
5    SEQ  EFDLVMVFWELLVLLPHVALELQFVGDLPPESDEQHFTLQRDLSLEVSVRPGSGISARPS
SEG  .....xxxxxxxxxxxxxxxx.....
PRD  cccchhhhhhhhhchhhhhhhhhhhccccccccchhhhhhhccccceeecccccccccccc
MEM  ...MMMMMMMMMMMMMMMM.....
10   SEQ  SGTKEKGGRRDLQIKVSARPYHLFQGPKPDLVIGFNSGFALKDTWLRSLPRLQSLRVPAF
SEG  .....
PRD  cccccccccceeeeeecccccccccccccccccccccccccccccccccccccccccccc
MEM  .....
15   SEQ  FTESSEYSCVMDGQTMAVATGGGTSPPAQPNPFRSPFRLRAADNCMSWYCNAFIFHLVYKP
SEG  .....
PRD  cccccceeeeeeccccceeeeeeccccccccccccccccccccchhhhhcchhhhhhhhhhhhhhhccc
MEM  .....
20   SEQ  AQGSGARPAAGPPPPSPTPSAPPAPTRRRRGEKKPGRGARRRK
SEG  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
PRD  cccccccccccccccccccccccccccccccccchhhhhcccccccccccccccccc
MEM  .....
25   SEQ
SEG
PRD
MEM

```

(No Prosite data available for DKFZphtes3_7n12.1)

(No Pfam data available for DKFZphtes3_7n12.1)

DKFZphtes3_9e1b

35 group: transmembrane protein

DKFZphtes3_9e1b encodes a novel 539 amino acid protein without similarity to known proteins.

40 The novel protein contains 1 transmembrane region. The only EST described so far is from testis.
No informative BLAST results; No predictive prosite, pfam or SCOP motife.

45 The new protein can find application in studying the expression profile of testis-specific genes and as a new marker for testicular cells.

50 putative protein

1 EST hit
perhaps complete cds.

55 Sequenced by DKFZ

Locus: unknown

Insert length: 2011 bp

Poly A stretch at pos. 1986, no polyadenylation signal found

```
5      1 CATGGCAACA TGAGCAGTGC TGAGATAATT GGTCTACAA ATCTTATAAT
      51 TCTGCTAGAG GATGAAGTCT TTGCCGATTT TTTCAACACA TTTCTTTCCC
     101 TCCCGGTTTT TGGTCAGACA CCATTTTATA CTGTTGAAAA TTCACAGTGG
     151 AGCTTGTGGC CAGAAATACC TTGTAACCTG ATTGCCAAAT ACAAAGGGTT
     201 ATTGACCTGG TTGGAAAAAT GCCGATTACC TTTCTTCTGT AAAACAAACT
    10 251 TGTGTTTCCA TTACATTCTC TGTCAGGAGT TCATCAGTTT CATTAAGTCC
     301 CCAGAAGGAG CCAAGATGAT GAGATGGAAA AAGGCAGACC AGTGGCTACT
     351 CCAGAAATGC ATTGGCGGGG TCAGAGGGAT GTGGCGCTTC TATTCCTACC
     401 TCACAGGCAG TGCAGGTGAA GAATTGGTGG ATTTCTGGAT CCTGCTGAG
     451 AACATCCTGA GCATAGATGA GATGGACCTG GAAGTGAGAG ACTACTACCT
    15 501 GTCCCTCCTC CTCATGCTGA GGGCCACTCA TCTGCAGGAG GGCTCCAGGG
     551 TGGTAACCCT CTGTAACATG AACATCAAGT CCCTCCTGAA CCTCTCCATC
     601 TGGCATCCCA ACCAATCAAC CACTAGGAGG GAGATCCTGA GCCACATGCA
     651 GAAAGTGGCT CTGTTCAAAC TCCAGAGCTA TTGGCTTCCC AACTTTTACA
     701 CCCACACCAA GATGACCATG GCCAAGGAGG AAGCATGCCA TGGTCTGATG
    20 751 CAAGAGTACG AGACTCGCTT ATACAGCGTT TGCTACACCC ACATAGGAGG
     801 GCTCCCTCTG AACATGAGCA TCAAGAAGTG CCACCACTTT CAGAAAACGGT
     851 ACTCAAGCAG GAAAGCCAAG AGGAAGATGT GGCAATTGGT AGATCCTGAC
     901 TCTTGGTCTC TGGAAATGGA TCTCAAGCCA GATGCTATTG GTATGCCCCT
     951 ACAGGAGACA TGTCCTCAAG AGAAGGTGGT TATACAAATG CCTTCCCTGA
    25 1001 AAATGGCTTC TTCAAAGGAA ACAAGAATCA GTTCCCTGGA AAAGGATATG
     1051 CATTATGCAA AAATATCCAG CATGGAGAAT AAAGCCAAGA GCCACCTCCA
     1101 CATGGAAGCC CCCTTTGAGA CAAAGGTCTC TACCCACCTG AGGACTGTCA
     1151 TCCCATTGT CAATCACTCC TCCAAGATGA CAATTCAGAA GGCCATCAAG
     1201 CAAAGCTTCT CTTAGGATA CATCCACTTG GCCTTGTTGT CTGATGCCTG
    30 1251 TGCAGGGAAC CCTTTCCGGG ACCACCTGAA GAAGCTGAAT TTGAAAGTGG
     1301 AGATCCAACT TCTTGACCTC TGGCAGGACT TGCAGCATTT CCTCAGTGTC
     1351 CTTCTGAATA AAAAAAGAA TGGGAATGCA ATCTTTCGTC ACTTGCTGGG
     1401 TGACAGAATC TGCGAGCTCT ACCTGAATGA GCAGATTGGT CCGTGCTTAC
     1451 CACTCAAATC CCAAACCATT CAGGGCCTGA AGGAACTATT GCCCTCTGGG
    35 1501 GATGTGATCC CCTGGATTCC CAAAGCCCAG AAGGAGATTG GCAAGATGCT
     1551 CAGTCCCTGG TATGATGAGT TTCTAGATGA AGAGGACTAC TGGTTTCTCC
     1601 TTTTACGGT AGGAAGGACT TTGGGTTAGG AAGGAATCAT GAGGATGAGG
     1651 GAAGAAGAAA GAGTAATTAC TGTTTTAAAA GGGTTATGTG TTAAAGTAAA
     1701 TGAAATTGTT ATTTTTCCTA GAGTCAACCA AAGATCAGCA TGGTCCCTGT
    40 1751 TGTTCTAAAG CTAAACCTCT CAAGGAAAAG GACTCAGTGC ATAAGATGAC
     1801 TTTGGTGAAA CCCCCTCTCT ACTAAAAATA CAAAAAATTA GCCGGGCGTA
     1851 GTGGCGGGCG CCTGTAGTCC CAGCTACTTG GGAGGCTGAG GCAGGAGAAT
     1901 GGTGTGAACC CGGGAGGCGG AGCTTGCAGT GAGCCGAGAT CCCGCCACTG
     1951 CACGCCAGCC TGGGCGACAG AGCGAGACTC CGTCTCAAAA AAAAAAAAAA
    45 2001 AAAAAAAAAA G
```

BLAST Results

50

No BLAST result

Medline entries

55

No Medline entry

Peptide information for frame 1

5

ORF from 10 bp to 1626 bp; peptide length: 539

Category: putative protein

Classification: no clue

```

10      1 MSSAEIIGST NLIILLEDEV FADFFNTFLS LPVFGQTPFY TVENSQWSLW
      51 PEIPCNIKAK YKGLLTWLEK CRLPFFCKTN LCFHYILCQE FISFIKSPEG
     101 AKMMRWKKAD QWLLQKCIGG VRGMWRFYSY LTGSAGEELV DFWILAENIL
     151 SIDEMDLEVR DYYLSLLLML RATHLQEGSR VVTLCNMNIK SLLNLSIWHP
     201 NQSTTRREIL SHMQKVALFK LQSYWLPNFY THTKMTMAKE EACHGLMQEY
     15  251 ETRLYSVCYT HIGGLPLNMS IKKCHHFQKR YSSRKAKRKM WQLVDPDSSWS
     301 LEMDLKPDAT GMPLQETCPQ EKVVIMPSL KMASSKETRI SSLEKDMHYA
     351 KISSMENKAK SHLHMEAPFE TKVSTHLRTV IPIVNHSSKM TIQKAIKQSF
     401 SLGYIHLALC ADACAGNPFR DHLKKLNLKV EIQLLDLWQD LQHFSLVLLN
     451 NKKNGNAIFR HLLGDRICEL YLNEQIGPCL PLKSQTIQGL KELLPSGDVI
    20  501 PWIPKAQKEI CKMLSPWYDE FLDEEDYWFL LFTVGRTLG

```

BLASTP hits

25

No BLASTP hits available

Alert BLASTP hits for DKFZphtes3_9e1b, frame 1

30. No Alert BLASTP hits found

Pedant information for DKFZphtes3_9e1b, frame 1

35

Report for DKFZphtes3_9e1b.1

```

[LENGTH] 542
[MW]      62906.06
40 [pI]      8.35
    [KW]      Alpha_Beta

```

```

45  SEQ  HGNMSSAEIIGSTNLIILLEDEVFADFFNTFLSLPVFGQTPFYTVENSQWSLWPEIPCNI
    PRD  cccccceeeccccceehhhhhhhhhccccccccccccccccccccccccccccccccch

    SEQ  IAKYKGLLTWLEKCRLPFFCKTNLCFHYILCQEFISFIKSPEGAKMMRWKKADQWLLQKC
    PRD  hhhccccceccccccccccccceehhhhhhhhhhhccccchhhhhhhcchhhhhhhhh

50  SEQ  IGGVRGMWRFYSYLTGSAGEELVDFWILAENILSIDEMDLEVRDYYLSLLLMLRATHLQE
    PRD  cccccceeeccccccccccccchhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhhhccc

    SEQ  GSRVVTLCNMNIKSLNLSIWHPNQSTTRREILSHMQKVALFKLQSYWLPNFYTHTKMTM
    PRD  cceeeccccchhhhhhhhhhhccccccccchhhhhhhhhhhhhhhhhhhhhccccccchhhhhhh

55  SEQ  AKEEACHGLMQEYETRLYSVCYTHIGGLPLNMSIKKCHHFQKRYSSRKAKRKMWQLVDPD
    PRD  hhhhhhhhhhhhhhhhhheeeccccccccccccccccccccchhhhhhhhhhhhhheeeccc

```


CONSENSUS: C-x-[DN]-x(4)-[FY]-x-C-x-C.

NAME: Vitamin K-dependent carboxylation domain.

5 CONSENSUS: x(12)-E-x(3)-E-x-C-x(6)-[DEN]-x-[LIVMFY]-x(9)-[FYW].

NAME: Phosphopantetheine attachment site.

CONSENSUS: [DEQGSTALMKRH]-[LIVMFYSTAC]-[GNQ]-[LIVMFYAG]-[DNEKHS]-S-[LIVMST]-

10 CONSENSUS: {PCFY}-[STAGCPQLIVMF]-[LIVMATN]-[DENQGTAKRHLN]-[LIVMWSTA]-[LIVGSTACR]-

CONSENSUS: x(2)-[LIVMFA].

15 NAME: Acyl carrier protein phosphopantetheine domain profile.

NAME: Prokaryotic membrane lipoprotein lipid attachment site.

20 CONSENSUS: {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGQ]-[AGS]-C.

NAME: Prokaryotic N-terminal methylation site.

CONSENSUS: [KRHEQSTAG]-G-[FYLIVM]-[EST]-[LT]-[LIVP]-E-[LIVMFWSTAG](14).

25 NAME: Prenyl group binding site (CAAX box).

CONSENSUS: C-[DENQ]-[LIVM]-x>.

NAME: Protein splicing signature.

30 CONSENSUS: [DNEG]-x-[LIVFA]-[LIVMY]-[LVAST]-H-N-[STC].

NAME: Endoplasmic reticulum targeting sequence.

CONSENSUS: [KRHQSA]-[DENQ]-E-L>.

NAME: Microbodies C-terminal targeting signal.

35 CONSENSUS: [STAGCN]-[ERKH]-[LIVMAFY]>.

NAME: Gram-positive cocci surface proteins 'anchoring' hexapeptide.

40 CONSENSUS: L-P-x-T-G-[STGAVDE].

NAME: Bipartite nuclear targeting sequence.

NAME: Cell attachment sequence.

45 CONSENSUS: R-G-D.

NAME: ATP/GTP-binding site motif A (P-loop).

CONSENSUS: [AG]-x(4)-G-K-[ST].

NAME: Cyclic nucleotide-binding domain signature 1.

50 CONSENSUS: [LIVM]-[VIC]-x(2)-G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x(2)-G.

NAME: Cyclic nucleotide-binding domain signature 2.

55 CONSENSUS: [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].

NAME: cAMP/cGMP binding motif.

- NAME: EF-hand calcium-binding domain.
 CONSENSUS: D-x-[DNS]-[ILVFYW]-[DENSTG]-[DNQGRK]-[GP]-
 [LIVMC]-[DENQSTAGC]-x(2)-
 CONSENSUS: [DE]-[LIVMFYW].
- 5 NAME: Actinin-type actin-binding domain signature 1.
 CONSENSUS: [EQ]-x(2)-[ATV]-[FY]-x(2)-W-x-N.
- 10 NAME: Actinin-type actin-binding domain signature 2.
 CONSENSUS: [LIVM]-x-[SGN]-[LIVM]-[DAGHE]-[SAG]-x-[DNEAG]-
 [LIVM]-x-[DEAG]-x(4)-
 CONSENSUS: [LIVM]-x-[LM]-[SAG]-[LIVM]-[LIVMT]-W-x-[LIVM](2).
- 15 NAME: Anaphylatoxin domain signature.
 CONSENSUS: [CSH]-C-x(2)-[GAP]-x(7,8)-[GASTDEQR]-C-[GASTDEQL]-
 x(3,9)-[GASTDEQN]-x(2)-
 CONSENSUS: [CE]-x(6,7)-C-C.
- 20 NAME: Anaphylatoxin domain profile.
- NAME: Apple domain.
 CONSENSUS: C-x(3)-[LIVMFY]-x(5)-[LIVMFY]-x(3)-[DENQ]-
 [LIVMFY]-x(10)-C-x(3)-C-T-
 CONSENSUS: x(4)-C-x-[LIVMFY]-F-x-[FY]-x(13,14)-C-x-[LIVMFY]-
 25 [RK]-x-[ST]-x(14,15)-
 CONSENSUS: S-G-x-[ST]-[LIVMFY]-x(2)-C.
- NAME: Band 4.1 family domain signature 1.
 CONSENSUS: W-[LIV]-x(3)-[KRQ]-x-[LIVM]-x(2)-[QH]-x(0,2)-
 30 [LIVMF]-x(6,8)-[LIVMF]-
 CONSENSUS: x(3,5)-F-[FY]-x(2)-[DENS].
- NAME: Band 4.1 family domain signature 2.
 CONSENSUS: [HYW]-x(9)-[DENQSTV]-[SA]-x(3)-[FY]-[LIVM]-x(2)-
 35 [ACV]-x(2)-[LM]-x(2)-
 CONSENSUS: [FY]-G-x-[DENQST]-[LIVMFYS].
- NAME: Band 4.1 family domain profile.
- 40 NAME: Clq domain signature.
 CONSENSUS: F-x(5)-[ND]-x(4)-[FYWL]-x(6)-F-x(5)-G-x-Y-x-F-x-
 [FY].
- NAME: C-terminal cystine knot signature.
 45 CONSENSUS: C-C-x(13)-C-x(2)-[GN]-x(12)-C-x-C-x(2,4)-C.
- NAME: C-terminal cystine knot profile.
- NAME: CUB domain profile.
 50 NAME: Death domain profile.
- NAME: EGF-like domain signature 1.
 CONSENSUS: C-x-C-x(5)-G-x(2)-C.
- 55 NAME: EGF-like domain signature 2.
 CONSENSUS: C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C.

- NAME: Calcium-binding EGF-like domain pattern signature.
 CONSENSUS: [DEQN]-x-[DEQN](2)-C-x(3,14)-C-x(3,7)-C-x-[DN]-x(4)-[FY]-x-C.
- 5 NAME: Laminin-type EGF-like (LE) domain signature.
 CONSENSUS: C-x(1,2)-C-x(5)-G-x(2)-C-x(2)-C-x(3,4)-[FYW]-x(3,15)-C.
- 10 NAME: Coagulation factors 5/8 type C domain (FA58C) signature 1.
 CONSENSUS: [GAS]-W-x(7,15)-[FYW]-[LIV]-x-[LIVFA]-[GSTDEN]-x(6)-[LIVF]-x(2)-[IV]-x-
 CONSENSUS: [LIVT]-[QKM]-G.
- 15 NAME: Coagulation factors 5/8 type C domain (FA58C) signature 2.
 CONSENSUS: P-x(8,10)-[LM]-R-x-[GE]-[LIVP]-x-G-C.
- 20 NAME: Forkhead-associated (FHA) domain profile.
 NAME: Fibrinogen beta and gamma chains C-terminal domain signature.
 CONSENSUS: W-W-[LIVMFYW]-x(2)-C-x(2)-[GSA]-x(2)-N-G.
- 25 NAME: Type I fibronectin domain.
 CONSENSUS: C-x(6,8)-[LFY]-x(5)-[FYW]-x-[RK]-x(8,10)-C-x-C-x(6,9)-C.
- 30 NAME: Type II fibronectin collagen-binding domain.
 CONSENSUS: C-x(2)-P-F-x-[FYWI]-x(7)-C-x(8,10)-W-C-x(4)-[DNSR]-[FYW]-x(3,5)-[FYW]-x-
 CONSENSUS: [FYWI]-C.
- 35 NAME: Hemopexin domain signature.
 CONSENSUS: [LIFAT]-x(3)-W-x(2,3)-[PE]-x(2)-[LIVMFY]-[DENQS]-[STA]-[AV]-[LIVMFY].
- 40 NAME: Kringle domain signature.
 CONSENSUS: [FY]-C-R-N-P-[DNR].
 NAME: Kringle domain profile.
- 45 NAME: LDL-receptor class A (LDLRA) domain signature.
 CONSENSUS: C-[VILMA]-x(5)-C-[DNH]-x(3)-[DENQHT]-C-x(3,4)-[STADE]-[DEH]-[DE]-x(1,5)-
 CONSENSUS: C.
- NAME: LDL-receptor class A (LDLRA) domain profile.
- 50 NAME: C-type lectin domain signature.
 CONSENSUS: C-[LIVMFYATG]-x(5,12)-[WL]-x-[DNSR]-x(2)-C-x(5,6)-[FYWLIVSTA]-[LIVMSTA]-
 CONSENSUS: C.
- 55 NAME: C-type lectin domain profile.
 NAME: Link domain signature.
 CONSENSUS: C-x(15)-A-x(3,4)-G-x(3)-C-x(2)-G-x(8,9)-P-x(7)-C.

- NAME: Osteonectin domain signature 1.
 5 CONSENSUS: C-x-[DN]-x(2)-C-x(2)-G-[KRH]-x-C-x(6,7)-P-x-C-x-C-x(3,5)-C-P.
- NAME: Osteonectin domain signature 2.
 CONSENSUS: F-P-x-R-[IM]-x-D-W-L-x-[NQ].
- NAME: Somatomedin B domain signature.
 10 CONSENSUS: C-x-C-x(3)-C-x(5)-C-C-x-[DN]-[FY]-x(3)-C.
- NAME: Thyroglobulin type-1 repeat signature.
 CONSENSUS: [FYWHP]-x-P-x-C-x(3,4)-G-x-[FYW]-x(3)-Q-C-x(4,10)-C-[FYW]-C-V-x(3,4)-
 15 CONSENSUS: [SG].
- NAME: P-type 'Trefoil' domain signature.
 CONSENSUS: R-x(2)-C-x-[FYPT]-x(3,4)-[ST]-x(3)-C-x(4)-C-C-[FYWH].
 20
- NAME: Cellulose-binding domain, bacterial type.
 CONSENSUS: W-N-[STAGR]-[STDN]-[LIVM]-x(2)-[GST]-x-[GST]-x(2)-[LIVMT]-[GA].
- NAME: Cellulose-binding domain, fungal type.
 25 CONSENSUS: C-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM]-x(2)-Q-C.
- NAME: Chitin recognition or binding domain signature.
 30 CONSENSUS: C-x(4,5)-C-C-S-x(2)-G-x-C-G-x(4)-[FYW]-C.
- NAME: Barwin domain signature 1.
 CONSENSUS: C-G-[KR]-C-L-x-V-x-N.
- NAME: Barwin domain signature 2.
 35 CONSENSUS: V-[DN]-Y-[EQ]-F-V-[DN]-C.
- NAME: BIR repeat.
 CONSENSUS: [HKEPILVY]-x(2)-R-x(3,7)-[FYW]-x(11,14)-[STAN]-G-[LMF]-X-[FYHDA]-X(4)-
 40 CONSENSUS: [DESL]-X(2,3)-C-X(2)-C-X(6)-[WA]-X(9)-H-X(4)-[PRSD]-X-C-X(2)-[LIVMA].
- NAME: WAP-type 'four-disulfide core' domain signature.
 45 CONSENSUS: C-x-{C}-[DN]-x(2)-C-x(5)-C-C.
- NAME: Phorbol esters / diacylglycerol binding domain.
 CONSENSUS: H-x-[LIVMFYW]-x(8,11)-C-x(2)-C-x(3)-[LIVMFC]-x(5,10)-C-x(2)-C-x(4)-[HD]-
 50 CONSENSUS: x(2)-C-x(5,9)-C.
- NAME: C2 domain signature.
 CONSENSUS: [ACG]-x(2)-L-x(2,3)-D-x(1,2)-[NGSTLIF]-[GTMR]-x-[STAP]-D-[PA]-[FY].
 55
- NAME: C2-domain profile.
- NAME: CAP-Gly domain signature.

CONSENSUS: G-x(8,10)-[FYW]-x-G-[LIVM]-x-[LIVMFY]-x(4)-G-K-
 [NH]-x-G-[STAR]-x(2)-G-
 CONSENSUS: x(2)-[LY]-F.

- 5 NAME: Ly-b / u-PAR domain signature.
 CONSENSUS: [EQR]-C-[LIVMFYAH]-x-C-x(5,8)-C-x(3,8)-[EDNQSTV]-
 C-[C]-x(5)-C-
 CONSENSUS: x(12,24)-C.
- 10 NAME: MAM domain signature.
 CONSENSUS: G-x-[LIVMFY](2)-x(3)-[STA]-x(10,11)-[LV]-x(4)-
 [LIVMF]-x(6,7)-C-[LIVM]-x-
 CONSENSUS: F-x-[LIVMFY]-x(3)-[GSC].
- 15 NAME: MAM domain profile.
 NAME: PH domain profile.
- 20 NAME: Phosphotyrosine interaction domain (PID) profile.
 NAME: Src homology 2 (SH2) domain profile.
 NAME: Src homology 3 (SH3) domain profile.
- 25 NAME: VWFC domain signature.
 CONSENSUS: C-x(2,3)-C-x-C-x(6,14)-C-x(3,4)-C-x(2,10)-C-
 x(9,16)-C-C-x(2,4)-C.
- 30 NAME: WW/rsp5/WWP domain signature.
 CONSENSUS: W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-
 [FYW]-x(2)-P.
 NAME: WW/rsp5/WWP domain profile.
- 35 NAME: ZP domain signature.
 CONSENSUS: [LIVMFYW]-x(7)-[STAPDNL]-x(3)-[LIVMFYW]-x-
 [LIVMFYW]-x-[LIVMFYW]-x(2)-C-
 CONSENSUS: [LIVMFYW]-x-[ST]-[PSL]-x(2,4)-[DENS]-x-[STADNQLF]-
 x(6)-[LIVM](2)-x(3,4)-
 40 CONSENSUS: C.
- NAME: S-layer homology domain signature.
 CONSENSUS: [LVFYT]-x-[DA]-x(2,5)-[DNGSATPHY]-[WYFPDA]-x(4)-
 [LIV]-x(2)-[GTALV]-
 45 CONSENSUS: x(4,6)-[LIVFYC]-x(2)-G-x-[PGSTA]-x(2,3)-[MFYA]-x-
 [PGAV]-x(3,10)-[LIVMA]-
 CONSENSUS: [STKR]-[RY]-x-[EQ]-x-[STALIVM].
- NAME: 'Homeobox' domain signature.
 50 CONSENSUS: [LIVMFYGG]-[ASLVR]-x(2)-[LIVMSTACN]-x-[LIVM]-x(4)-
 [LIV]-[RKNESTAIY]-
 CONSENSUS: [LIVFSTNKH]-W-[FYVC]-x-[ENDQTAH]-x(5)-[RKNAIMW].
- NAME: 'Homeobox' domain profile.
- 55 NAME: 'Homeobox' antennapedia-type protein signature.
 CONSENSUS: [LIVMFE]-[FY]-P-W-M-[KRQTA].

- NAME: 'Homeobox' engrailed-type protein signature.
 CONSENSUS: L-M-A-Q-G-L-Y-N.
- 5 NAME: 'Paired box' domain signature.
 CONSENSUS: R-P-C-x(11)-C-V-S.
- NAME: 'POU' domain signature 1.
 CONSENSUS: [RKQ]-R-[LIM]-x-[LF]-G-[LIVMFY]-x-Q-x-[DNQ]-V-G.
- 10 NAME: 'POU' domain signature 2.
 CONSENSUS: S-Q-[ST]-[TA]-I-[SC]-R-F-E-x-[LSQ]-x-[LI]-[ST].
- NAME: Zinc finger, C2H2 type, domain.
 CONSENSUS: C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.
- 15 NAME: Zinc finger, C3HC4 type (RING finger), signature.
 CONSENSUS: C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA].
- NAME: Nuclear hormones receptors DNA-binding region
 signature.
 CONSENSUS: C-x(2)-C-x-[DE]-x(5)-[HN]-[FY]-x(4)-C-x(2)-C-x(2)-
 F-F-x-R.
- 20 NAME: GATA-type zinc finger domain.
 CONSENSUS: C-x-[DN]-C-x(4,5)-[ST]-x(2)-W-[HR]-[RK]-x(3)-[GN]-
 x(3,4)-C-N-[AS]-C.
- NAME: Poly(ADP-ribose) polymerase zinc finger domain
 signature.
 CONSENSUS: C-[KR]-x-C-x(3)-I-x-K-x(3)-[RG]-x(16,18)-W-[FYH]-
 H-x(2)-C.
- 30 NAME: Poly(ADP-ribose) polymerase zinc finger domain
 profile.
- 35 NAME: Fungal Zn(2)-Cys(6) binuclear cluster domain
 signature.
 CONSENSUS: [GASTPV]-C-x(2)-C-[RKHSTACW]-x(2)-[RKHQ]-x(2)-C-
 x(5,12)-C-x(2)-C-x(6,8)-
- 40 CONSENSUS: C.
- NAME: Fungal Zn(2)-Cys(6) binuclear cluster domain profile.
- NAME: Prokaryotic dksA/traR C4-type zinc finger.
 CONSENSUS: C-[DES]-x-C-x(3)-I-x(3)-R-x(4)-P-x(4)-C-x(2)-C.
- 45 NAME: Copper-fist domain signature.
 CONSENSUS: M-[LIVMF](3)-x(3)-K-[MY]-A-C-x(2)-C-I-[KR]-x-H-
 [KR]-x(3)-C-x-H-x(8)-
- 50 CONSENSUS: [KR]-x-[KR]-G-R-P.
- NAME: Copper fist DNA binding domain profile.
- NAME: Leucine zipper pattern.
 CONSENSUS: L-x(6)-L-x(6)-L-x(6)-L.
- 55 NAME: bZIP transcription factors basic domain signature.

- CONSENSUS: [KR]-x(1,3)-[RKSAQ]-N-x(2)-[SAQ](2)-x-[RKTAENQ]-x-
 R-x-[RK].
- 5 NAME: Myb DNA-binding domain repeat signature 1.
 CONSENSUS: W-[ET]-x(2)-E-[DE]-x(2)-[LIV].
- NAME: Myb DNA-binding domain repeat signature 2.
 CONSENSUS: W-x(2)-[LI]-[SAG]-x(4,5)-R-x(8)-[YW]-x(3)-[LIVM].
- 10 NAME: Myc-type, 'helix-loop-helix' dimerization domain
 signature.
 CONSENSUS: [DENSTAP]-K-[LIVMWAGSN]-[FYWCPHKR]-[LIVT]-[LIV]-
 x(2)-[STAV]-[LIVMSTAC]-x-
 CONSENSUS: [VMFYH]-[LIVMTA]-[P]-[P]-[LIVMSR].
- 15 NAME: p53 tumor antigen signature.
 CONSENSUS: M-C-N-S-S-C-M-G-G-M-N-R-R.
- 20 NAME: CBF-A/NF-YB subunit signature.
 CONSENSUS: C-V-S-E-x-I-S-F-[LIVM]-T-[SG]-E-A-[SC]-[DE]-[KRQ]-
 C.
- NAME: CBF-B/NF-YA subunit signature.
 CONSENSUS: Y-V-N-A-K-Q-Y-x-R-I-L-K-R-R-x-A-R-A-K-L-E.
- 25 NAME: 'Cold-shock' DNA-binding domain signature.
 CONSENSUS: [FY]-G-F-I-x(6,7)-[DER]-[LIVM]-F-x-H-x-[STKR]-x-
 [LIVMFY].
- 30 NAME: CTF/NF-I signature.
 CONSENSUS: R-K-R-K-Y-F-K-K-H-E-K-R.
- NAME: Ets-domain signature 1.
 CONSENSUS: L-[FYW]-[QEDH]-F-[LI]-[LVQK]-x-[LI]-L.
- 35 NAME: Ets-domain signature 2.
 CONSENSUS: [RKH]-x(2)-M-x-Y-[DENQ]-x-[LIVM]-[STAG]-R-[STAG]-
 [LI]-R-x-Y.
- 40 NAME: Ets-domain profile.
- NAME: Fork head domain signature 1.
 CONSENSUS: [KR]-P-[PTQ]-[FYLVQH]-S-[FY]-x(2)-[LIVM]-x(3,4)-
 [AC]-[LIM].
- 45 NAME: Fork head domain signature 2.
 CONSENSUS: W-[QKR]-[NS]-S-[LIV]-R-H.
- NAME: Fork head domain profile.
- 50 NAME: HSF-type DNA-binding domain signature.
 CONSENSUS: L-x(3)-[FY]-K-H-x-N-x-[STAN]-S-F-[LIVM]-R-Q-L-
 [NH]-x-Y-x-[FYW]-[RKH]-K-
 CONSENSUS: [LIVM].
- 55 NAME: Tryptophan pentad repeat (IRF family) signature.
 CONSENSUS: W-x-[DNH]-x(5)-[LIVF]-x-[IV]-P-W-x-H-x(9,10)-[DE]-
 x(2)-[LIVF]-F-[KRQ]-x-

CONSENSUS: [WR]-A.

NAME: LIM domain signature.

5 CONSENSUS: C-x(2)-C-x(15,21)-[FYWH]-H-x(2)-[CH]-x(2)-C-x(2)-
C-x(3)-[LIVMF].

NAME: LIM domain profile.

NAME: NF-kappa-B/Rel/dorsal domain signature.

10 CONSENSUS: F-R-Y-x-C-E-G.

NAME: MADS-box domain signature.

CONSENSUS: R-x-[RK]-x(5)-I-x-[DN]-x(3)-[KR]-x(2)-T-[FY]-x-
[RK](3)-x(2)-[LIVM]-x-

15 CONSENSUS: K(2)-A-x-E-[LIVM]-[ST]-x-L-x(4)-[LIVM]-x-
[LIVM](3)-x(6)-[LIVMF]-x(2)-
CONSENSUS: [FY].

NAME: MADS-box domain profile.

20

NAME: T-box domain signature 1.

CONSENSUS: L-W-x(2)-[FC]-x(3,4)-[NT]-E-M-[LIV](2)-T-x(2)-G-
[RG]-[KRQ].

NAME: T-box domain signature 2.

25 CONSENSUS: [LIVMYW]-H-[PADH]-[DEN]-[GS]-x(3)-G-x(2)-W-M-x(3)-
[IVA]-x-F.

NAME: TEA domain signature.

30 CONSENSUS: G-R-N-E-L-I-x(2)-Y-I-x(3)-[TC]-x(3)-R-T-[RK](2)-Q-
[LIVM]-S-S-H-[LIVM]-
CONSENSUS: Q-V.

NAME: Transcription factor TFIIB repeat signature.

35 CONSENSUS: G-[KR]-x(3)-[STAGN]-x-[LIVMYA]-[GSTA](2)-[CSAV]-
[LIVM]-[LIVMFY]-[LIVMA]-
CONSENSUS: [GSA]-[STAC].

NAME: Transcription factor TFIID repeat signature.

40 CONSENSUS: Y-x-P-x(2)-[IF]-x(2)-[LIVM](2)-x-[KRH]-x(3)-P-
[RKQ]-x(3)-L-[LIVM]-F-x-
CONSENSUS: [STN]-G-[KR]-[LIVM]-x(3)-G-[TAGL]-[KR]-x(7)-[AGC]-
x(7)-[LIVM].

NAME: TFIIS zinc ribbon domain signature.

45 CONSENSUS: C-x(2)-C-x(9)-[LIVMQSAR]-[QH]-[STQL]-[RA]-[SACR]-
x-[DE]-[DET]-[PGSEA]-
CONSENSUS: x(6)-C-x(2,5)-C-x(3)-[FW].

NAME: TSC-22 / dip / bun family signature.

50 CONSENSUS: M-D-L-V-K-x-H-L-x(2)-A-V-R-E-E-V-E.

NAME: Prokaryotic transcription elongation factors signature 1.

55 CONSENSUS: [ST]-x(2)-[GS]-x(3)-[LI]-x(2)-E-L-x(2)-L-x(3,4)-R-
x(2)-[IV]-x(3)-[LIV]-
CONSENSUS: x(6)-G-D-x(2)-E-N-[GSA]-x-Y.

- NAME: Prokaryotic transcription elongation factors signature 2.
 CONSENSUS: S-x(2)-S-P-[LIVM]-[AG]-x-[SAG]-[LIVM]-[LIVMY]-x(4)-[DG]-[DE].
- 5 NAME: DEAD-box subfamily ATP-dependent helicases signature.
 CONSENSUS: [LIVMF](2)-D-E-A-D-[RKEN]-x-[LIVMFYGSTN].
- 10 NAME: DEAH-box subfamily ATP-dependent helicases signature.
 CONSENSUS: [GSAH]-x-[LIVMF](3)-D-E-[ALIV]-H-[NECR].
- NAME: Eukaryotic putative RNA-binding region RNP-1 signature.
 CONSENSUS: [RK]-G-[EDRKHPCG]-[AGSCI]-[FY]-[LIVA]-x-[FYLM].
- 15 NAME: Fibrillarin signature.
 CONSENSUS: [GST]-[LIVMAP]-V-Y-A-[IV]-E-[FY]-[SA]-x-R-x(2)-R-[DE].
- 20 NAME: MCM family signature.
 CONSENSUS: G-[IVT]-[LVAC](2)-[IVT]-D-[DE]-[FL]-[DNST].
- NAME: MCM family domain.
- 25 NAME: XPA protein signature 1.
 CONSENSUS: C-x-[DE]-C-x(3)-[LIVMF]-x(1,2)-D-x(2)-L-x(3)-F-x(4)-C-x(2)-C.
- 30 NAME: XPA protein signature 2.
 CONSENSUS: [LIVM](2)-T-[KR]-T-E-x-K-x-[DE]-Y-[LIVMF](2)-x-D-x-[DE].
- NAME: XPG protein signature 1.
 CONSENSUS: [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K.
- 35 NAME: XPG protein signature 2.
 CONSENSUS: [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]-[QS]-[CLM].
- 40 NAME: Bacterial regulatory proteins, araC family signature.
 CONSENSUS: [KRQ]-[LIVMA]-x(2)-[GSTALIV]-[FYWPGDN]-x(2)-[LIVMSA]-x(4,9)-[LIVMF]-
 CONSENSUS: x(2)-[LIVMSTA]-[GSTACIL]-x(3)-[GANQRF]-[LIVMFY]-
 45 x(4,5)-[LFY]-x(3)-
 CONSENSUS: [FYIVA]-[FYWHCM]-x(3)-[GSADENQKR]-x-[NSTAPKL]-[PARL].
- NAME: Bacterial regulatory proteins, araC family DNA-binding domain profile.
- 50 NAME: Bacterial regulatory proteins, arsR family signature.
 CONSENSUS: C-x(2)-D-[LIVM]-x(6)-[ST]-x(4)-S-[HYR]-[HQ].
- 55 NAME: Bacterial regulatory proteins, asnC family signature.
 CONSENSUS: [GSTAP]-x(2)-[DNEA]-[LIVM]-[GSA]-x(2)-[LIVMFY]-[GN]-[LIVMST]-[ST]-x(6)-R-
 CONSENSUS: [LVT]-x(2)-[LIVM]-x(3)-G.

- NAME: Bacterial regulatory proteins, crp family signature.
 CONSENSUS: [LIVM]-[STAG]-[RHNV]-x(2)-[LIM]-[GA]-x-[LIVMFYA]-
 [LIVSC]-[GA]-x-[STACN]-
 5 CONSENSUS: x(2)-[MST]-x-[GSTN]-R-x-[LIVMF]-x(2)-[LIVMF].
- NAME: Bacterial regulatory proteins, deoR family signature.
 CONSENSUS: R-x(3)-[LIVM]-x(3)-[LIVM]-x(16,17)-[STA]-x(2)-T-
 [LIVMA]-[RH]-[KRNA]-D-
 10 CONSENSUS: [LIVMF].
- NAME: Bacterial regulatory proteins, gntR family signature.
 CONSENSUS: [LIVAPKR]-[PILV]-x-[EQTIVMR]-x(2)-[LIVM]-x(3)-
 [LIVMFYK]-x-[LIVFT]-
 15 CONSENSUS: [DNGSTK]-[RGTLV]-x-[STAIVP]-[LIVA]-x(2)-[STAGV]-
 [LIVMFYH]-x(2)-[LMA].
- NAME: Bacterial regulatory proteins, iclR family signature.
 CONSENSUS: [GA]-x(3)-[DS]-x(2)-E-x(6)-[CSA]-[LIVM]-[GSA]-
 20 x(2)-[LIVM]-[FYH]-[DN].
- NAME: Bacterial regulatory proteins, lacI family signature.
 CONSENSUS: [LIVM]-x-[DE]-[LIVM]-A-x(2)-[STAGV]-x-V-[GSTP]-
 x(2)-[STAG]-[LIVMA]-x(2)-
 25 CONSENSUS: [LIVMFYAN]-[LIVMC].
- NAME: Bacterial regulatory proteins, luxR family signature.
 CONSENSUS: [GDC]-x(2)-[NSTAVY]-x(2)-[IV]-[GSTA]-x(2)-
 [LIVMFYWCT]-x-[LIVMFYWCR]-x(3)-
 30 CONSENSUS: [NST]-[LIVM]-x(5)-[NRHSA]-[LIVMSTA]-x(2)-[KR].
- NAME: Bacterial regulatory proteins, lysR family signature.
 CONSENSUS: [NQKRHSTAG]-[LIVMFYTA]-x(2)-[STAGLV]-[STAG]-x(4)-
 [LIVMYCTQR]-[PSTANLVER]-
 35 CONSENSUS: x-[PSTAGQV]-[PSTAGNVMF]-[LIVMFA]-[STAGH]-x(2)-
 [LIVMF]-x(2)-[LIVMFW]-
 CONSENSUS: [RKEAV]-x(2)-[LIVMFYNTAE]-x(3)-[LIMVT].
- NAME: Bacterial regulatory proteins, marR family signature.
 CONSENSUS: [STNA]-[LIA]-x-[RNGS]-x(4)-[LM]-[EIV]-x(2)-[GES]-
 [LFYW]-[LIVC]-x(7)-
 40 CONSENSUS: [DN]-[RKQG]-[RK]-x(6)-T-x(2)-[GA].
- NAME: Bacterial regulatory proteins, merR family signature.
 CONSENSUS: [GSA]-x-[LIVMFA]-[ASM]-x(2)-[STACLIV]-[GSDENQR]-
 [LIVC]-[STANHK]-x(3)-
 45 CONSENSUS: [LIVM]-[RHF]-x-[YW]-[DEQ]-x(2,3)-[GHDNQ]-
 [LIVMF](2).
- NAME: Bacterial regulatory proteins, tetR family signature.
 CONSENSUS: G-[LIVMFYS]-x(2,3)-[TS]-[LIVMT]-x(2)-[LIVM]-x(5)-
 [LIVQS]-[STAGENQH]-x-
 50 CONSENSUS: [GPAR]-x-[LIVMF]-[FYST]-x-[HFY]-[FV]-x-[DNST]-K-
 x(2)-[LIVM].
- NAME: Transcriptional antiterminators bglG family signature.
 CONSENSUS: [ST]-x-H-x(2)-[FA](2)-[LIVM]-[EQK]-R-x(2)-[QNK].

- NAME: Sigma-54 factors family signature 1.
 CONSENSUS: P-[LIVM]-x-[LIVM]-x(2)-[LIVM]-A-x(2)-[LIVMF]-x(2)-[HS]-x-S-T-[LIVM]-S-R.
- 5 NAME: Sigma-54 factors family signature 2.
 CONSENSUS: R-R-T-[IV]-[AT]-K-Y-R.
- NAME: Sigma-54 factors family profile.
- 10 NAME: Sigma-70 factors family signature 1.
 CONSENSUS: [DE]-[LIVMF](2)-[HEQS]-x-G-x-[LIVMFA]-G-L-[LIVMFYE]-x-[GSAM]-[LIVMAP].
- NAME: Sigma-70 factors family signature 2.
 15 CONSENSUS: [STN]-x(2)-[DEQ]-[LIVM]-[GAS]-x(4)-[LIVMF]-[PSTG]-x(3)-[LIVMA]-x-[NQR]-
 CONSENSUS: [LIVMA]-[EQH]-x(3)-[LIVMFW]-x(2)-[LIVM].
- NAME: Sigma-70 factors ECF subfamily signature.
 20 CONSENSUS: [STAIV]-[PQDEL]-[DE]-[LIV]-[LIVTA]-Q-x-[STAV]-[LIVMFYC]-[LIVMAK]-x-
 CONSENSUS: [GSTAIV]-[LIMFYWQ]-x(12,14)-[STAP]-[FYW]-[LIF]-x(2)-[IV].
- 25 NAME: Sigma-54 interaction domain ATP-binding region A signature.
 CONSENSUS: [LIVMFY](3)-x-G-[DEQ]-[STE]-G-[STAV]-G-K-x(2)-[LIVMFY].
- 30 NAME: Sigma-54 interaction domain ATP-binding region B signature.
 CONSENSUS: [GS]-x-[LIVMF]-x(2)-A-[DNEQASH]-[GNEK]-G-[STIM]-[LIVMFY](3)-[DE]-[EK]-
 CONSENSUS: [LIVM].
- 35 NAME: Sigma-54 interaction domain C-terminal part signature.
 CONSENSUS: [FYW]-P-[GS]-N-[LIVM]-R-[EQ]-L-x-[NHAT].
- NAME: Sigma-54 interaction domain profile.
- 40 NAME: Single-strand binding protein family signature 1.
 CONSENSUS: [LIVMF]-[NST]-[KRT]-[LIVM]-x-[LIVMF](2)-G-[NHRK]-[LIVM]-[GST]-x-[DET].
- 45 NAME: Single-strand binding protein family signature 2.
 CONSENSUS: T-x-W-[HY]-[RNS]-[LIVM]-x-[LIVMF]-[FY]-[NGKR].
- NAME: Bacterial histone-like DNA-binding proteins signature.
 50 CONSENSUS: [GSK]-F-x(2)-[LIVMF]-x(4)-[RKEQA]-x(2)-[RST]-x-[GA]-x-[KN]-P-x-T.
- NAME: Dps protein family signature 1.
 CONSENSUS: H-[FW]-x-[LIVM]-x-G-x(5)-[LV]-H-x(3)-[DE].
- 55 NAME: Dps protein family signature 2.
 CONSENSUS: [LIVMFY]-[DH]-x-[LIVM]-[GA]-E-R-x(3)-[LIF]-[GDN]-x(2)-[PA].

NAME: DNA repair protein radC family signature.
 CONSENSUS: H-N-H-P-S-G.

NAME: recA signature.
 5 CONSENSUS: A-L-[KR]-[IF]-[FY]-[STA]-[STAD]-[LIVM]-R.

NAME: RecF protein signature 1.
 CONSENSUS: P-[ED]-x(3)-[LIVM](2)-x-G-[GSA]-P-x(2)-R-R-x-[FY]-[LIVM]-D.

10 NAME: RecF protein signature 2.
 CONSENSUS: [LIVMFY](2)-x-D-x(2,3)-[SA]-[EH]-L-D-x(2)-[KRH]-x(3)-L.

15 NAME: RecR protein signature.
 CONSENSUS: C-x(2)-C-x(3)-[ST]-x(4)-C-x-I-C-x(4)-R.

NAME: Histone H2A signature.
 20 CONSENSUS: [AC]-G-L-x-F-P-V.

NAME: Histone H2B signature.
 CONSENSUS: [KR]-E-[LIVM]-[EQ]-T-x(2)-[KR]-x-[LIVM](2)-x-[PAG]-[DE]-L-x-[KR]-H-A-
 CONSENSUS: [LIVM]-[STA]-E-G.

25 NAME: Histone H3 signature 1.
 CONSENSUS: K-A-P-R-K-Q-L.

NAME: Histone H3 signature 2.
 30 CONSENSUS: P-F-x-[RA]-L-[VA]-[KRQ]-[DEG]-[IV].

NAME: Histone H4 signature.
 CONSENSUS: G-A-K-R-H.

35 NAME: HMGI/2 signature.
 CONSENSUS: [FI]-S-[KR]-K-C-S-[EK]-R-W-K-T-M.

NAME: HMG-I and HMG-Y DNA-binding domain (A+T-hook).
 CONSENSUS: [AT]-x(1,2)-[RK](2)-[GP]-R-G-R-P-[RK]-x.

40 NAME: HMGI4 and HMGI7 signature.
 CONSENSUS: R-R-S-A-R-L-S-A-[RK]-P.

NAME: Bromodomain signature.
 45 CONSENSUS: [STANVF]-x(2)-F-x(4)-[DNS]-x(5,7)-[DENQTF]-Y-[HFY]-x(2)-[LIVMFY]-x(3)-
 CONSENSUS: [LIVM]-x(4)-[LIVM]-x(6,8)-Y-x(12,13)-[LIVM]-x(2)-N-[SACF]-x(2)-[FY].

50 NAME: Bromodomain profile.

NAME: Chromo domain signature.
 CONSENSUS: [FYL]-x-[LIVMC]-[KR]-W-x-[GDNR]-[FYWLE]-x(5,6)-[ST]-W-[ES]-[PSTDN]-x(3)-
 55 CONSENSUS: [LIVMC].

NAME: Chromo and chromo shadow domain profile.

- NAME: Regulator of chromosome condensation (RCC1) signature 1.
 1. CONSENSUS: G-x-N-D-x(2)-[AV]-L-G-R-x-T.
- 5 NAME: Regulator of chromosome condensation (RCC1) signature 2.
 CONSENSUS: [LIVMFA]-[STAGC](2)-G-x(2)-H-[STAGLI]-[LIVMFA]-x-[LIVM].
- 10 NAME: Protamine P1 signature.
 CONSENSUS: [AV]-R-[NFY]-R-x(2,3)-[ST]-x-S-x-S.
- NAME: Nuclear transition protein 1 signature.
 CONSENSUS: S-K-R-K-Y-R-K.
- 15 NAME: Nuclear transition protein 2 signature 1.
 CONSENSUS: H-x(3)-H-S-[NS]-S-x-P-Q-S.
- NAME: Nuclear transition protein 2 signature 2.
 20 CONSENSUS: K-x-R-K-x(2)-E-G-K-x(2)-K-[KR]-K.
- NAME: Ribosomal protein L1 signature.
 CONSENSUS: [IM]-x(2)-[LIVA]-x(2,3)-[LIVM]-G-x(2)-[LMS]-[GSNH]-[PTKR]-[KRAV]-G-x-
 25 CONSENSUS: [LMF]-P-[DENSTK].
- NAME: Ribosomal protein L2 signature.
 CONSENSUS: P-x(2)-R-G-[STAIV](2)-x-N-[APK]-x-[DE].
- 30 NAME: Ribosomal protein L3 signature.
 CONSENSUS: [FL]-x(6)-[DN]-x(2)-[AGS]-x-[ST]-x-G-[KRH]-G-x(2)-G-x(3)-R.
- NAME: Ribosomal protein L5 signature.
 35 CONSENSUS: [LIVM]-x(2)-[LIVM]-[STAC]-[GE]-[QV]-x(2)-[LIVMA]-x-[STC]-x-[STAG]-[KR]-
 CONSENSUS: x-[STA].
- NAME: Ribosomal protein L6 signature 1.
 40 CONSENSUS: [PS]-[DENS]-x-Y-K-[GA]-K-G-[LIVM].
- NAME: Ribosomal protein L6 signature 2.
 CONSENSUS: Q-x(3)-[LIVM]-x(2)-[KR]-x(2)-R-x-F-x-D-G-[LIVM]-Y-[LIVM]-x(2)-[KR].
- 45 NAME: Ribosomal protein L9 signature.
 CONSENSUS: G-x(2)-[GN]-x(4)-V-x(2)-G-[FY]-x(2)-N-[FY]-L-x(5)-[GA]-x(3)-[STN].
- 50 NAME: Ribosomal protein L10 signature.
 CONSENSUS: [DEH]-x(2)-[GS]-[LIVMF]-[STN]-[VA]-x-[DEQK]-[LIVMA]-x(2)-[LIM]-R.
- NAME: Ribosomal protein L11 signature.
 55 CONSENSUS: [RKN]-x-[LIVM]-x-G-[ST]-x(2)-[SNQ]-[LIVM]-G-x(2)-[LIVM]-x(0,1)-[DENG].
- NAME: Ribosomal protein L13 signature.

CONSENSUS: [LIVM]-[KRV]-[GK]-M-[LIV]-[PS]-x(4,5)-[GS]-
[NQEKRA]-x(5)-[LIVM]-x-[AIV]-
CONSENSUS: [LFY]-x-[GDN].

5 NAME: Ribosomal protein L14 signature.
CONSENSUS: [GA]-[LIV](3)-x(9,10)-[DNS]-G-x(4)-[FY]-x(2)-[ENT]-
x(2)-V-[LIV].

10 NAME: Ribosomal protein L15 signature.
CONSENSUS: K-[LIVM](2)-[GAL]-x-[GT]-x-[LIVMA]-x(2,5)-[LIVM]-
x-[LIVMF]-x(3,4)-
CONSENSUS: [LIVMFC]-[EST]-x(2)-A-x(3)-[LIVM]-x(3)-G.

15 NAME: Ribosomal protein L1b signature 1.
CONSENSUS: [KR]-R-x-[GSAC]-[KQVA]-[LIVM]-W-[LIVM]-[KR]-
[LIVM]-[LFY]-[AP].

20 NAME: Ribosomal protein L1b signature 2.
CONSENSUS: R-M-G-x-[GR]-K-G-x(4)-[FWKR].

NAME: Ribosomal protein L17 signature.
CONSENSUS: I-x-[ST]-[GT]-x(2)-[KR]-x-K-x(6)-[DE]-x-[LIMV]-
[LIVMT]-T-x-[STAG]-[KR].

25 NAME: Ribosomal protein L19 signature.
CONSENSUS: [RT]-[KRSVY]-[GSA]-x-V-[RS]-[KR]-[SA]-K-L-Y-Y-L-R.

30 NAME: Ribosomal protein L20 signature.
CONSENSUS: K-x(3)-[KRC]-x-[LIVM]-W-[IV]-[STNALV]-R-[LIVM]-N-
x(3)-[RKH].

35 NAME: Ribosomal protein L21 signature.
CONSENSUS: [IVT]-x(3)-[KR]-x(3)-[KRQ]-K-x(6)-G-[HF]-R-[RQ]-
x(2)-T.

NAME: Ribosomal protein L22 signature.
CONSENSUS: [RKQN]-x(4)-[RH]-[GAS]-x-G-[KRQS]-x(9)-[HDN]-
[LIVM]-x-[LIVMS]-x-[LIVM].

40 NAME: Ribosomal protein L23 signature.
CONSENSUS: [RK](2)-[AM]-[IVFYT]-[IV]-[RKT]-L-[STANQK]-x(7)-
[LIVMFT].

45 NAME: Ribosomal protein L24 signature.
CONSENSUS: [GDEN]-D-x-V-x-[IV]-[LIVMA]-x-G-x(2)-[KA]-[GN]-
x(2,3)-[GA]-x-[IV].

50 NAME: Ribosomal protein L27 signature.
CONSENSUS: G-x-[LIVM](2)-x-R-Q-R-G-x(5)-G.

NAME: Ribosomal protein L29 signature.
CONSENSUS: [KNQS]-[PSTL]-x(2)-[LIMFA]-[KRGSA]-x-[LIVYSTA]-
[KR]-[KRH]-[DESTANRL]-
CONSENSUS: [LIV]-A-[KRCQVT]-[LIVMA].

55 NAME: Ribosomal protein L30 signature.
CONSENSUS: [IVT]-[LIVM]-x(2)-[LF]-x-[LI]-x-[KRHQEG]-x(2)-
[STNQH]-x-[IVT].

- CONSENSUS: x(10)-[LMS]-[LIV]-x(2)-[LIVA]-x(2)-[LMFY]-[IVT].
- NAME: Ribosomal protein L31 signature.
 5 CONSENSUS: H-P-F-[FY]-[TI]-x(9)-G-R-[AV]-x-[KR].
- NAME: Ribosomal protein L33 signature.
 CONSENSUS: Y-x-[ST]-x-[KR]-[NS]-x(4)-[PAT]-x(1,2)-[LIVM]-[EA]-x(2)-K-[FY]-[CS].
- 10 NAME: Ribosomal protein L34 signature.
 CONSENSUS: K-[RG]-T-[FYWL]-[EQS]-x(5)-[KRHS]-x(4,5)-G-F-x(2)-R.
- NAME: Ribosomal protein L35 signature.
 15 CONSENSUS: [LIVM]-K-[TV]-x(2)-[GSA]-[SAIL]-x-K-R-[LIVMFY]-[KRL].
- NAME: Ribosomal protein L36 signature.
 CONSENSUS: C-x(2)-C-x(2)-[LIVM]-x-R-x(3)-[LIVMN]-x-[LIVM]-x-C-x(3,4)-[KR]-H-x-Q-x-Q.
- 20 NAME: Ribosomal protein L1e signature.
 CONSENSUS: N-x(3)-[KR]-x(2)-A-[LIVT]-x-S-A-[LIV]-x-A-[ST]-[SGA]-x(7)-[RK]-G-H.
- 25 NAME: Ribosomal protein L1e signature.
 CONSENSUS: N-x(2)-P-L-R-R-x(4)-[FY]-V-I-A-T-S-x-K.
- NAME: Ribosomal protein L7Ae signature.
 30 CONSENSUS: [CA]-x(4)-[IV]-P-[FY]-x(2)-[LIVM]-x-[GSQ]-[KRQ]-x(2)-L-G.
- NAME: Ribosomal protein L10e signature.
 CONSENSUS: R-x-A-[FYW]-G-K-[PA]-x-G-x(2)-A-R-V.
- 35 NAME: Ribosomal protein L13e signature.
 CONSENSUS: [KR]-Y-x(2)-K-[LIVM]-R-[STA]-G-[KR]-G-F-[ST]-L-x-E.
- NAME: Ribosomal protein L15e signature.
 40 CONSENSUS: [DE]-[KR]-A-R-x-L-G-[FY]-x-[SAP]-x(2)-G-[LIVMFY](4)-R-x-R-V-x-R-G.
- NAME: Ribosomal protein L18e signature.
 45 CONSENSUS: [KRE]-x-L-x(2)-[PS]-[KR]-x(2)-[RH]-[PSA]-x-[LIVM]-[NS]-[LIVM]-x-[RK]-[LIVM].
- NAME: Ribosomal protein L19e signature.
 50 CONSENSUS: R-x-[KR]-x(5)-[KR]-x(3)-[KRH]-x(2)-G-x-G-x-R-x-G-x(3)-A-R-x(3)-[KQ]-[LIVM].
- NAME: Ribosomal protein L21e signature.
 55 CONSENSUS: G-[DE]-x-V-x(10)-[GV]-x(2)-[FYH]-x(2)-[FY]-x-G-x-T-G.
- NAME: Ribosomal protein L24e signature.

CONSENSUS: [FY]-x-[GS]-x(2)-[IV]-x-P-G-x-G-x(2)-[FYV]-x-
[KRHE]-x-D.

5 NAME: Ribosomal protein L27e signature.
CONSENSUS: G-K-N-x-W-F-F-x-K-L-R-F>.

NAME: Ribosomal protein L30e signature 1.
CONSENSUS: [STA]-x(5)-G-x-[QKR]-x(2)-[LIVM]-[KQT]-x(2)-[KR]-
x-G-x(2)-K-x-[LIVM](3).

10 NAME: Ribosomal protein L30e signature 2.
CONSENSUS: [DE]-L-G-[STA]-x(2)-G-[KR]-x(6)-[LIVM]-x-[LIVM]-x-
[DEN]-x-G.

15 NAME: Ribosomal protein L31e signature.
CONSENSUS: V-[KR]-[LIVM]-x(3)-[LIVM]-N-x-[AK]-x-W-x-[KR]-G.

NAME: Ribosomal protein L32e signature.
20 CONSENSUS: F-x-R-x(4)-[KR]-x(2)-[KR]-[LIVM]-x(3)-W-R-[KR]-
x(2)-G.

NAME: Ribosomal protein L34e signature.
CONSENSUS: Y-x-[ST]-x-S-[NY]-x(5)-[KR]-T-P-G.

25 NAME: Ribosomal protein L35Ae signature.
CONSENSUS: G-K-[LIVM]-x-R-x-H-G-x(2)-G-x-V-x-A-x-F-x(3)-[LI]-
P.

NAME: Ribosomal protein L36e signature.
30 CONSENSUS: P-Y-E-[KR]-R-x-[LIVM]-[DE]-[LIVM](2)-[KR].

NAME: Ribosomal protein L37e signature.
CONSENSUS: G-T-x-[SA]-x-G-x-[KR]-x(3)-[ST]-x(0,1)-H-x(2)-C-x-
R-C-G.

35 NAME: Ribosomal protein L39e signature.
CONSENSUS: [KRA]-T-x(3)-[LIVM]-[KRQF]-x-[NHS]-x(3)-R-[NHY]-W-
R-R.

40 NAME: Ribosomal protein L44e signature.
CONSENSUS: K-x-[TV]-K-K-x(2)-L-[KR]-x(2)-C.

NAME: Ribosomal protein S2 signature 1.
45 CONSENSUS: [LIVMFA]-x(2)-[LIVMFYC](2)-x-[STAC]-[GSTANQEK]-
[STALV]-[HY]-[LIVMF]-G.

NAME: Ribosomal protein S2 signature 2.
CONSENSUS: P-x(2)-[LIVMF](2)-[LIVMS]-x-[GDN]-x(3)-[DENL]-
x(3)-[LIVM]-x-E-x(4)-

50 CONSENSUS: [GNQKRH]-[LIVM]-[AP].

NAME: Ribosomal protein S3 signature.
CONSENSUS: [GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1,3)-
[LIVFCA]-x(3)-[LIV]-

55 CONSENSUS: [DENQ]-x(7)-[LMT]-x(2)-G-x(2)-G.

NAME: Ribosomal protein S4 signature.

CONSENSUS: [LIVM]-[DE]-x-R-L-x(3)-[LIVMC]-[VMFYHQ]-[KRT]-
x(3)-[STAGCF]-x-[ST]-x(3)-
CONSENSUS: [SAI]-[KR]-x-[LIVMF](2).

- 5 NAME: Ribosomal protein S5 signature.
CONSENSUS: G-[KRQ]-x(3)-[FY]-x-[ACV]-x(2)-[LIVMA]-[LIVM]-
[AG]-[DN]-x(2)-G-x-
CONSENSUS: [LIVM]-G-x-[SAG]-x(5,6)-[DEQ]-[LIVM]-x(2)-A-
[LIVMF].
- 10 NAME: Ribosomal protein S6 signature.
CONSENSUS: G-x-[KRC]-[DENQRH]-L-[SA]-Y-x-I-[KRNSA].
- 15 NAME: Ribosomal protein S7 signature.
CONSENSUS: [DENSK]-x-[LIVMET]-x(3)-[LIVMFT](2)-x(6)-G-K-[KR]-
x(5)-[LIVMF]-[LIVMFC]-
CONSENSUS: x(2)-[STA].
- 20 NAME: Ribosomal protein S8 signature.
CONSENSUS: [GE]-x(2)-[LIV](2)-[STY]-T-x(2)-G-[LIVM](2)-x(4)-
[AG]-[KRHAYI].
- 25 NAME: Ribosomal protein S9 signature.
CONSENSUS: G-G-G-x(2)-[GSA]-Q-x(2)-[SA]-x(3)-[GSA]-x-[GSTAV]-
[KR]-[GSAL]-[LIF].
- 30 NAME: Ribosomal protein S10 signature.
CONSENSUS: [AV]-x(3)-[GDNSR]-[LIVMSTA]-x(3)-G-P-[LIVM]-x-
[LIVM]-P-T.
- 35 NAME: Ribosomal protein S11 signature.
CONSENSUS: [LIVMF]-x-[GSTAC]-[LIVMF]-x(2)-[GSTAL]-x(0,1)-
[GSN]-[LIVMF]-x-[LIVM]-
CONSENSUS: x(4)-[DEN]-x-T-P-x-[PA]-[STCH]-[DN].
- 40 NAME: Ribosomal protein S12 signature.
CONSENSUS: [RK]-x-P-N-S-[AR]-x-R.
- 45 NAME: Ribosomal protein S13 signature.
CONSENSUS: [KRQS]-G-x-R-H-x(2)-[GSNH]-x(2)-[LIVMC]-R-G-Q.
- 50 NAME: Ribosomal protein S14 signature.
CONSENSUS: [RP]-x(0,1)-C-x(11,12)-[LIVMF]-x-[LIVMF]-[ESC]-
[RG]-x(3)-[RN].
- 55 NAME: Ribosomal protein S15 signature.
CONSENSUS: [LIVM]-x(2)-H-[LIVMFY]-x(5)-D-x(2)-[SAGN]-x(3)-
[LF]-x(9)-[LIVM]-x(2)-
CONSENSUS: [FY].
- NAME: Ribosomal protein S16 signature.
CONSENSUS: [LIVMT]-x-[LIVM]-[KR]-L-[STAK]-R-x-G-[AKR].
- NAME: Ribosomal protein S17 signature.
CONSENSUS: G-D-x-[LIV]-x-[LIVA]-x-[QEK]-x-[RK]-P-[LIV]-S.
- NAME: Ribosomal protein S18 signature.

CONSENSUS: [IV]-[DY]-Y-x(2)-[LIVMT]-x(2)-[LIVM]-x(2)-[FYT]-
[LIVM]-[EST]-[DERP]-x-
CONSENSUS: [GY]-K-[LIVM]-x(3)-R-[LIVMAS].

- 5 NAME: Ribosomal protein S19 signature.
CONSENSUS: [STDNQ]-G-[KRQM]-x(6)-[LIVM]-x(4)-[LIVM]-[GSD]-
x(2)-[LF]-[GAS]-[DE]-F-
CONSENSUS: x(2)-[ST].
- 10 NAME: Ribosomal protein S21 signature.
CONSENSUS: [DE]-x-A-[LY]-[KR]-R-F-K-[KR]-x(3)-[KR].
- NAME: Ribosomal protein S3Ae signature.
CONSENSUS: [LIV]-x-[GH]-R-[IV]-x-E-x-[SC]-L-x-D-L.
- 15 NAME: Ribosomal protein S4e signature.
CONSENSUS: H-x-K-R-[LIVM]-[SAN]-x-P-x(2)-W-x-[LIVM]-x-[KR].
- NAME: Ribosomal protein S6e signature.
20 CONSENSUS: [LIVM]-[STAMR]-G-G-x-D-x(2)-G-x-P-M.
- NAME: Ribosomal protein S7e signature.
CONSENSUS: [KR]-L-x-R-E-L-E-K-K-F-[SAP]-x-[KR]-H.
- 25 NAME: Ribosomal protein S8e signature.
CONSENSUS: R-x(2)-T-G-[GA]-x(5)-[HR]-K-[KR]-x-K-x-E-[LM]-G.
- NAME: Ribosomal protein S12e signature.
CONSENSUS: A-L-[KRQP]-x-V-L-x(2)-[SA]-x(3)-[DN]-G-L.
- 30 NAME: Ribosomal protein S17e signature.
CONSENSUS: A-x-I-x-[ST]-K-x-L-R-N-[KR]-I-A-G-[FY]-x-T-H.
- NAME: Ribosomal protein S19e signature.
35 CONSENSUS: P-x(6)-[SAN]-x(2)-[LIVMA]-x-R-x-[ALIV]-[LV]-Q-x-L-
[EQ].
- NAME: Ribosomal protein S21e signature.
CONSENSUS: L-Y-V-P-R-K-C-S-[SA].
- 40 NAME: Ribosomal protein S24e signature.
CONSENSUS: [FA]-G-x(2)-[KR]-[STA]-x-G-[FY]-[GA]-x-[LIVM]-Y-
[DN]-[SN].
- 45 NAME: Ribosomal protein S26e signature.
CONSENSUS: [YH]-C-V-S-C-A-I-H.
- NAME: Ribosomal protein S27e signature.
CONSENSUS: [QK]-C-x(2)-C-x(6)-F-[GS]-x-[PSA]-x(5)-C-x(2)-C-
50 [GS]-x(2)-L-x(2)-P-x-G.
- NAME: Ribosomal protein S28e signature.
CONSENSUS: E-[ST]-E-R-E-A-R-x-L.
- 55 NAME: DNA mismatch repair proteins mutL / hexB / PMS1
signature.
CONSENSUS: G-F-R-G-E-A-L.

- NAME: DNA mismatch repair proteins mutS family signature.
 CONSENSUS: [EST]-[LIVM]-x-[LIVM]-x-D-E-[LIVMY]-[GCI]-[RKH]-G-[GST]-x(4)-G.
- 5 NAME: mutT domain signature.
 CONSENSUS: G-x(5)-E-x(4)-[STAGC]-[LIVMAC]-x-R-E-[LIVMFT]-x-E-E.
- 10 NAME: DnaA protein signature.
 CONSENSUS: I-[GA]-x(2)-[LIVMF]-[SGDNK]-x(0,1)-[KR]-x-H-[STP]-[STV]-[LIVM](2)-x-
 CONSENSUS: [SA]-x(2)-[KRE]-[LIVM].
- 15 NAME: Small, acid-soluble spore proteins, alpha/beta type, signature 1.
 CONSENSUS: K-x-E-[LIV]-A-x-[DE]-[LIVMF]-G-[LIVMF].
- 20 NAME: Small, acid-soluble spore proteins, alpha/beta type, signature 2.
 CONSENSUS: [KR]-[SAQ]-x-G-x-V-G-G-x-[LIVM]-x-[KR](2)-[LIVM](2).
- 25 NAME: Zinc-containing alcohol dehydrogenases signature.
 CONSENSUS: G-H-E-x(2)-G-x(5)-[GA]-x(2)-[IVSAC].
- 30 NAME: Quinone oxidoreductase / zeta-crystallin signature.
 CONSENSUS: [GSD]-[DEQH]-x(2)-L-x(3)-[SA](2)-G-G-x-G-x(4)-Q-x(2)-[KR].
- 35 NAME: Iron-containing alcohol dehydrogenases signature 1.
 CONSENSUS: [STALIV]-[LIVF]-x-[DE]-x(6,7)-P-x(4)-[ALIV]-x-[GST]-x(2)-D-[TAIVM]-
 CONSENSUS: [LIVMF]-x(4)-E.
- 40 NAME: Iron-containing alcohol dehydrogenases signature 2.
 CONSENSUS: [GSW]-x-[LIVTSACD]-[GH]-x(2)-[GSAE]-[GSHYQ]-x-[LIVTP]-[GAST]-[GAS]-x(3)-
 CONSENSUS: [LIVMT]-x-[HNS]-[GA]-x-[GTAC].
- 45 NAME: Short-chain dehydrogenases/reductases family signature.
 CONSENSUS: [LIVSPADNK]-x(12)-Y-[PSTAGNCV]-[STAGNQCIVM]-[STAGC]-K-[PC]-[SAGFR]-
 CONSENSUS: [LIVMSTAGD]-x(2)-[LIVMFYW]-x(3)-[LIVMFYWGAPTHQ]-[GSACQRHM].
- 50 NAME: Aldo/keto reductase family signature 1.
 CONSENSUS: G-[FY]-R-[HSAL]-[LIVMF]-D-[STAGC]-[AS]-x(5)-E-x(2)-[LIVM]-G.
- 55 NAME: Aldo/keto reductase family signature 2.
 CONSENSUS: [LIVMFY]-x(9)-[KREQ]-x-[LIVM]-G-[LIVM]-[SC]-N-[FY].
- NAME: Aldo/keto reductase family putative active site signature.
 CONSENSUS: [LIVM]-[PAIV]-[KR]-[ST]-x(4)-R-x(2)-[GSTAEQK]-[NSL]-x(2)-[LIVMFA].

NAME: Homoserine dehydrogenase signature.

CONSENSUS: A-x(3)-G-[LIVMFY]-[ESTAG]-x(2,3)-[DNS]-P-x(2)-D-[LIVM]-x-G-x-D-x(3)-K.

NAME: NAD-dependent glycerol-3-phosphate dehydrogenase signature.

CONSENSUS: G-[AT]-[LIVM]-K-[DN]-[LIVM](2)-A-x-[GA]-x-G-[LIVMF]-x-[DE]-G-[LIVM]-x-

CONSENSUS: [LIVMFYW]-G-x-N.

NAME: FAD-dependent glycerol-3-phosphate dehydrogenase signature 1.

CONSENSUS: [IV]-G-G-G-x(2)-G-[STACV]-G-x-A-x-D-x(3)-R-G.

NAME: FAD-dependent glycerol-3-phosphate dehydrogenase signature 2.

CONSENSUS: G-G-K-x(2)-[GSTE]-Y-R-x(2)-A.

NAME: Mannitol dehydrogenases signature.

CONSENSUS: [LIVMY]-x-[FS]-x(2)-[STAGCV]-x-V-D-R-[IV]-x-[PS].

NAME: Histidinol dehydrogenase signature.

CONSENSUS: I-D-x(2)-A-G-P-[ST]-E-[LIVS]-[LIVMA](3)-[AC]-x(3)-A-x(4)-[LIVM]-[AV]-

CONSENSUS: [SACL]-[DE]-[LIVMFC]-[LIVM]-[SA]-x(2)-E-H.

NAME: L-lactate dehydrogenase active site.

CONSENSUS: [LIVMA]-G-[EQ]-H-G-[DN]-[ST].

NAME: D-isomer specific 2-hydroxyacid dehydrogenases NAD-binding signature.

CONSENSUS: [LIVMA]-[AG]-[IVT]-[LIVMFY]-[AG]-x-G-[NHKRQGSAC]-[LIV]-G-x(13,14)-

CONSENSUS: [LIVfMT]-x(2)-[FYwCTH]-[DNSTK].

NAME: D-isomer specific 2-hydroxyacid dehydrogenases signature 2.

CONSENSUS: [LIVMFYWA]-[LIVFYWC]-x(2)-[SAC]-[DNQHR]-[IVFA]-[LIVF]-x-[LIVF]-[HNI]-x-

CONSENSUS: P-x(4)-[STN]-x(2)-[LIVMF]-x-[GSDN].

NAME: D-isomer specific 2-hydroxyacid dehydrogenases signature 3.

CONSENSUS: [LMFATC]-[KPR]-x-[GSTDN]-x-[LIVMFYWR]-[LIVMFYW](2)-N-x-[STAGC]-R-[GP]-x-

CONSENSUS: [LIVH]-[LIVMC]-[DNV].

NAME: 3-hydroxyisobutyrate dehydrogenase signature.

CONSENSUS: [LIVMFY](2)-G-L-G-x-[MQ]-G-x-[PGS]-[MA]-[SA].

NAME: Hydroxymethylglutaryl-coenzyme A reductases signature 1.

CONSENSUS: [RKH]-x(6)-D-x-M-G-x-N-x-[LIVMA].

NAME: Hydroxymethylglutaryl-coenzyme A reductases signature 2.

CONSENSUS: [LIVM]-G-x-[LIVM]-G-G-[AG]-T.

- NAME: Hydroxymethylglutaryl-coenzyme A reductases signature 3.
 5 CONSENSUS: A-[LIVM]-x-[STAN]-x(2)-[LI]-x-[KRNQ]-[GSA]-H-[LM]-x-[FYLH].
- NAME: Hydroxymethylglutaryl-coenzyme A reductases profile.
- NAME: 3-hydroxyacyl-CoA dehydrogenase signature.
 10 CONSENSUS: [DNE]-x(2)-[GA]-F-[LIVMFY]-x-[NT]-R-x(3)-[PA]-[LIVMFY](2)-x(5)-
 CONSENSUS: [LIVMFYCT]-[LIVMFY]-x(2)-[GV].
- NAME: Malate dehydrogenase active site signature.
 15 CONSENSUS: [LIVM]-T-[TRKMN]-L-D-x(2)-R-[STA]-x(3)-[LIVMFY].
- NAME: Malic enzymes signature.
 CONSENSUS: F-x-[DV]-D-x(2)-G-T-[GSA]-x-[IV]-x-[LIVMA]-[GAST](2)-[LIVMF](2).
 20
- NAME: Isocitrate and isopropylmalate dehydrogenases signature.
 CONSENSUS: [NS]-[LIMYT]-[FYDN]-G-[DNT]-[IMVY]-x-[STGDN]-[DN]-x(2)-[SGAP]-x(3,4)-G-
 25 CONSENSUS: [STG]-[LIVMPA]-G-[LIVMF].
- NAME: 6-phosphogluconate dehydrogenase signature.
 CONSENSUS: [LIVM]-x-D-x(2)-[GA]-[NQS]-K-G-T-G-x-W.
- NAME: Glucose-6-phosphate dehydrogenase active site.
 30 CONSENSUS: D-H-Y-L-G-K-[EQK].
- NAME: IMP dehydrogenase / GMP reductase signature.
 CONSENSUS: [LIVM]-[RK]-[LIVM]-G-[LIVM]-G-x-G-S-[LIVM]-C-x-T.
 35
- NAME: Bacterial quinoprotein dehydrogenases signature 1.
 CONSENSUS: [DEN]-W-x(3)-G-[RK]-x(6)-[FYW]-S-x(4)-[LIVM]-N-x(2)-N-V-x(2)-L-[RK].
- NAME: Bacterial quinoprotein dehydrogenases signature 2.
 CONSENSUS: W-x(4)-Y-D-x(3)-[DN]-[LIVMFY](4)-x(2)-G-x(2)-[STA]-P.
 40
- NAME: FMN-dependent alpha-hydroxy acid dehydrogenases active site.
 45 CONSENSUS: S-N-H-G-[AG]-R-Q.
- NAME: GMC oxidoreductases signature 1.
 CONSENSUS: [GA]-[RKN]-x-[LIV]-G(2)-[GST](2)-x-[LIVM]-N-x(3)-[FYWA]-x(2)-[PAG]-x(5)-
 50 CONSENSUS: [DNESH].
- NAME: GMC oxidoreductases signature 2.
 CONSENSUS: [GS]-[PSTA]-x(2)-[ST]-P-x-[LIVM](2)-x(2)-S-G-[LIVM]-G.
 55
- NAME: Eukaryotic molybdopterin oxidoreductases signature.

CONSENSUS: [GA]-x(3)-[KRNQHT]-x(11,14)-[LIVMFYWS]-x(8)-
 [LIVMF]-x-C-x(2)-[DEN]-R-
 CONSENSUS: x(2)-[DE].

- 5 NAME: Prokaryotic molybdopterin oxidoreductases signature 1.
 CONSENSUS: [STAN]-x-[CH]-x(2,3)-C-[STAG]-[GSTVMF]-x-C-x-
 [LIVMFYW]-x-[LIVMA]-x(3,4)-
 CONSENSUS: [DENQKHT].
- 10 NAME: Prokaryotic molybdopterin oxidoreductases signature 2.
 CONSENSUS: [STA]-x-[STAC](2)-x(2)-[STA]-D-[LIVMY](2)-L-P-x-
 [STAC](2)-x(2)-E.
- 15 NAME: Prokaryotic molybdopterin oxidoreductases signature 3.
 CONSENSUS: A-x(3)-[GDT]-I-x-[DNQTK]-x-[DEA]-x-[LIVM]-x-
 [LIVMC]-x-[NS]-x(2)-[GS]-
 CONSENSUS: x(5)-A-x-[LIVM]-[ST].
- 20 NAME: Aldehyde dehydrogenases glutamic acid active site.
 CONSENSUS: [LIVMFGA]-E-[LIMSTAC]-[GS]-G-[KNLM]-[SADN]-
 [TAPFV].
- 25 NAME: Aldehyde dehydrogenases cysteine active site.
 CONSENSUS: [FYLVA]-x(3)-G-[QE]-x-C-[LIVMGSTANC]-[AGCN]-x-
 [GSTADNEKR].
- 30 NAME: Aspartate-semialdehyde dehydrogenase signature.
 CONSENSUS: [LIVM]-[SADN]-x(2)-C-x-R-[LIVM]-x(4)-[GSC]-H-
 [STA].
- 35 NAME: Glyceraldehyde 3-phosphate dehydrogenase active site.
 CONSENSUS: [ASV]-S-C-[NT]-T-x(2)-[LIM].
- NAME: N-acetyl-gamma-glutamyl-phosphate reductase active
 site.
 CONSENSUS: [LIVM]-[GSA]-x-P-G-C-[FY]-[AVP]-T-[GA]-x(3)-
 [GTAC]-[LIVM]-x-P.
- 40 NAME: Gamma-glutamyl phosphate reductase signature.
 CONSENSUS: V-x(5)-A-[LIV]-x-H-I-x(2)-[HY]-[GS]-[ST]-x-H-[ST]-
 [DE]-x-I.
- 45 NAME: Dihydrodipicolinate reductase signature.
 CONSENSUS: E-[IV]-x-E-x-H-x(3)-K-x-D-x-P-S-G-T-A.
- NAME: Dihydroorotate dehydrogenase signature 1.
 CONSENSUS: [GS]-x(4)-[GK]-[STA]-[IVSTA]-[GT]-x(3)-[NQR]-x-G-
 [NH]-x(2)-P-[RT].
- 50 NAME: Dihydroorotate dehydrogenase signature 2.
 CONSENSUS: [LIV](2)-[GSA]-x-G-G-[IV]-x-[STGN]-x(3)-[ACV]-
 x(6)-G-A.
- 55 NAME: Coproporphyrinogen III oxidase signature.
 CONSENSUS: K-x-W-C-x(2)-[FYH](3)-[LIVM]-x-H-R-x-E-x-R-G-
 [LIVM]-G-G-[LIVM]-F-F-D.

- NAME: Fumarate reductase / succinate dehydrogenase FAD-binding site.
 CONSENSUS: R-[EST]-H-[EST]-x(2)-A-x-G-G.
- 5 NAME: Acyl-CoA dehydrogenases signature 1.
 CONSENSUS: [GAC]-[LIVM]-[EST]-E-x(2)-[GSAN]-G-[EST]-D-x(2)-[GSA].
- 10 NAME: Acyl-CoA dehydrogenases signature 2.
 CONSENSUS: [QDE]-x(2)-G-[GS]-x-G-[LIVMFY]-x(2)-[DEN]-x(4)-[KR]-x(3)-[DEN].
- NAME: Alanine dehydrogenase & pyridine nucleotide transhydrogenase signature 1.
 15 CONSENSUS: G-[LIVM]-P-x-E-x(3)-N-E-x(1,3)-R-V-A-x-[EST]-P-x-[GST]-V-x(2)-L-x-[KRH]-
 CONSENSUS: x-G.
- NAME: Alanine dehydrogenase & pyridine nucleotide transhydrogenase signature 2.
 20 CONSENSUS: [LIVM](2)-G-[GA]-G-x-A-G-x(2)-[SA]-x(3)-[GA]-x-[SG]-[LIVM]-G-A-x-V-
 CONSENSUS: x(3)-D.
- 25 NAME: Glu / Leu / Phe / Val dehydrogenases active site.
 CONSENSUS: [LIV]-x(2)-G-G-[SAG]-K-x-[GV]-x(3)-[DNST]-[EPL].
- NAME: D-amino acid oxidases signature.
 30 CONSENSUS: [LIVM](2)-H-[NHA]-Y-G-x-[GSA](2)-x-G-x(5)-G-x-A.
- NAME: Pyridoxamine 5'-phosphate oxidase signature.
 CONSENSUS: [LIVF]-E-F-W-[QHG]-x(4)-R-[LIVM]-H-[DNE]-R.
- NAME: Copper amine oxidase topaquinone signature.
 35 CONSENSUS: [LIVM]-[LIVMA]-[LIVM]-x(4)-T-x(2)-N-Y-[DE]-[YN].
- NAME: Copper amine oxidase copper-binding site signature.
 CONSENSUS: T-x-G-x(2)-H-[LIVMF]-x(3)-E-[DE]-x-P.
- 40 NAME: Lysyl oxidase putative copper-binding region signature.
 CONSENSUS: W-E-W-H-S-C-H-Q-H-Y-H.
- NAME: Delta 1-pyrroline-5-carboxylate reductase signature.
 45 CONSENSUS: [PALF]-x(2,3)-[LIV]-x(3)-[LIVM]-[ESTAC]-[STV]-x-[GAN]-G-x-T-x(2)-[AG]-
 CONSENSUS: [LIV]-x(2)-[LMF]-[DENQK].
- NAME: Dihydrofolate reductase signature.
 50 CONSENSUS: [LVAGC]-[LIF]-G-x(4)-[LIVMF]-P-W-x(4,5)-[DE]-x(3)-[FYIV]-x(3)-[STIQ].
- NAME: Tetrahydrofolate dehydrogenase/cyclohydrolase signature 1.
 55 CONSENSUS: [EQ]-x-[EQK]-[LIVM](2)-x(2)-[LIVM]-x(2)-[LIVMY]-N-x-[DN]-x(5)-[LIVMF](3)-
 CONSENSUS: Q-L-P-[LV].

NAME: Tetrahydrofolate dehydrogenase/cyclohydrolase

signature 2.

CONSENSUS: P-G-G-V-G-P-[MF]-T-[IV].

5 NAME: Oxygen oxidoreductases covalent FAD-binding site.
CONSENSUS: P-x(10)-[DE]-[LIVM]-x(3)-[LIVM]-x(9)-[LIVM]-x(3)-
[GSA]-[GST]-G-H.

10 NAME: Pyridine nucleotide-disulphide oxidoreductases class-I
active site.
CONSENSUS: G-G-x-C-[LIVA]-x(2)-G-C-[LIVM]-P.

NAME: Pyridine nucleotide-disulphide oxidoreductases class-II active site.
15 CONSENSUS: C-x(2)-C-D-[GA]-x(2,4)-[FY]-x(4)-[LIVM]-x-
[LIVM](2)-G(3)-[DN].

NAME: Respiratory-chain NADH dehydrogenase subunit 1
signature 1.
20 CONSENSUS: G-[LIVMFYKRS]-[LIVMAGP]-Q-x-[LIVMFY]-x-D-[AGIM]-
[LIVMFTA]-K-[LVMYST]-
CONSENSUS: [LIVMFYGI]-x-[KR]-[EQG].

NAME: Respiratory-chain NADH dehydrogenase subunit 1
signature 2.
25 CONSENSUS: P-F-D-[LIVMFYQ]-[STAGPVM]-E-[GAC]-E-x-[EQ]-
[LIVMS]-x(2)-G.

NAME: Respiratory-chain NADH dehydrogenase 20 Kd subunit
signature.
30 CONSENSUS: [GN]-x-D-[KRST]-[LIVMF](2)-P-[IV]-D-[LIVMFYW](2)-
x-P-x-C-P-[PT].

NAME: Respiratory-chain NADH dehydrogenase 24 Kd subunit
signature.
35 CONSENSUS: D-x(2)-F-[ST]-x(5)-C-L-G-x-C-x(2)-[GA]-P.

NAME: Respiratory chain NADH dehydrogenase 30 Kd subunit
signature.
40 CONSENSUS: E-R-E-x(2)-[DE]-[LIVMF](2)-x(6)-[HK]-x(3)-[KRP]-x-
[LIVM]-[LIVMS].

NAME: Respiratory chain NADH dehydrogenase 49 Kd subunit
signature.
45 CONSENSUS: [LIVMH]-H-[RT]-[GA]-x-E-K-[LIVMT]-x-E-x-[KRQ].

NAME: Respiratory-chain NADH dehydrogenase 51 Kd subunit
signature 1.
50 CONSENSUS: G-[AM]-G-[AR]-Y-[LIVM]-C-G-[DE](2)-[STA](2)-
[LIM](2)-[EN]-S.

NAME: Respiratory-chain NADH dehydrogenase 51 Kd subunit
signature 2.
55 CONSENSUS: E-S-C-G-x-C-x-P-C-R-x-G.

NAME: Respiratory-chain NADH dehydrogenase 75 Kd subunit
signature 1.
CONSENSUS: P-x(2)-C-[YWS]-x(7)-G-x-C-R-x-C.

- NAME: Respiratory-chain NADH dehydrogenase 75 Kd subunit signature 2.
 5 CONSENSUS: C-P-x-C-[DE]-x-[GS](2)-x-C-x-L-Q.
- NAME: Respiratory-chain NADH dehydrogenase 75 Kd subunit signature 3.
 CONSENSUS: R-C-[LIVM]-x-C-x-R-C-[LIVM]-x-[FY].
- 10 NAME: Nitrite and sulfite reductases iron-sulfur/siroheme-binding site.
 CONSENSUS: [STV]-G-C-x(3)-C-x(6)-[DE]-[LIVMF]-[GAT]-[LIVMF].
- NAME: Uricase signature.
 15 CONSENSUS: L-x-[LV]-L-K-[ST]-T-x-S-x-F-x(2)-[FY]-x(4)-[FY].
- NAME: Heme-copper oxidase catalytic subunit, copper B binding region signature.
 20 CONSENSUS: [YWG]-[LIVFYWTA](2)-[VGS]-H-[LNP]-x-V-x(44,47)-H-H.
- NAME: CO II and nitrous oxide reductase dinuclear copper centers signature.
 25 CONSENSUS: V-x-H-x(33,40)-C-x(3)-C-x(3)-H-x(2)-M.
- NAME: Cytochrome c oxidase subunit Vb, zinc binding region signature.
 CONSENSUS: [LIVM](2)-[FYW]-x(10)-C-x(2)-C-G-x(2)-[FY]-K-L.
- 30 NAME: Multicopper oxidases signature 1.
 CONSENSUS: G-x-[FYW]-x-[LIVMFYW]-x-[CST]-x(8)-G-[LM]-x(3)-[LIVMFYW].
- NAME: Multicopper oxidases signature 2.
 35 CONSENSUS: H-C-H-x(3)-H-x(3)-[AG]-[LM].
- NAME: Peroxidases proximal heme-ligand signature.
 40 CONSENSUS: [DET]-[LIVMTA]-x(2)-[LIVM]-[LIVMSTAG]-[SAG]-[LIVMSTAG]-H-[STA]-[LIVMFY].
- NAME: Peroxidases active site signature.
 CONSENSUS: [SGATV]-x(3)-[LIVMA]-R-[LIVMA]-x-[FW]-H-x-[SAC].
- NAME: Catalase proximal heme-ligand signature.
 45 CONSENSUS: R-[LIVMFSTAN]-F-[GASTNP]-Y-x-D-[AST]-[QEH].
- NAME: Catalase proximal active site signature.
 50 CONSENSUS: [IF]-x-[RH]-x(4)-[EQ]-R-x(2)-H-x(2)-[GAS]-[GASTF]-[GAST].
- NAME: Glutathione peroxidases selenocysteine active site.
 CONSENSUS: [GN]-[RKHNFC]-x-[LIVMFC]-[LIVMF](2)-x-N-[VT]-x-[STC]-x-C-[GA]-x-T.
- 55 NAME: Glutathione peroxidases signature 2.
 CONSENSUS: [LIV]-[AGD]-F-P-[CS]-[NG]-Q-F.
- NAME: Lipoxygenases iron-binding region signature 1.

CONSENSUS: H-[EQ]-x(3)-H-x-[LM]-[NQR]-[GST]-H-[LIVMSTAC](3)-E.

NAME: Lipoxygenases iron-binding region signature 2.

5 CONSENSUS: [LIVMA]-H-P-[LIVM]-x-[KRQ]-[LIVMF](2)-x-[AP]-H.

NAME: Extradial ring-cleavage dioxygenases signature.

CONSENSUS: [GNTIV]-x-H-x(5,7)-[LIVMF]-Y-x(2)-[DENTA]-P-x-[GP]-x(2,3)-E.

10

NAME: Intradiol ring-cleavage dioxygenases signature.

CONSENSUS: [LIVM]-x-G-x-[LIVM]-x(4)-[GS]-x(2)-[LIVM]-x(4)-[LIVM]-[DE]-[LIVMFY]-

CONSENSUS: x(6)-G-x-[FY].

15

NAME: Indoleamine 2,3-dioxygenase signature 1.

CONSENSUS: G-G-S-[AN]-[GA]-Q-S-S-x(2)-Q.

NAME: Indoleamine 2,3-dioxygenase signature 2.

20 CONSENSUS: [FY]-L-[DQ]-[DE]-[LIVM]-x(2)-Y-M-x(3)-H-[KR].

NAME: Bacterial ring hydroxylating dioxygenases alpha-subunit signature.

CONSENSUS: C-x-H-R-[GA]-x(8)-G-N-x(5)-C-x-[FY]-H.

25

NAME: Bacterial luciferase subunits signature.

CONSENSUS: [GA]-[LIVM]-P-[LIVM]-x-[LIVMFY]-x-W-x(6)-[RK]-x(6)-Y-x(3)-[AR].

30

NAME: ubiH/C04b monooxygenase family signature.

CONSENSUS: H-P-[LIV]-[AG]-G-Q-G-x-N-x-G-x(2)-D.

NAME: Biopterin-dependent aromatic amino acid hydroxylases signature.

35 CONSENSUS: P-D-x(2)-H-[DE]-[LI]-[LIVMF]-G-H-[LIVMC]-P.

NAME: Copper type II, ascorbate-dependent monooxygenases signature 1.

CONSENSUS: H-H-M-x(2)-F-x-C.

40

NAME: Copper type II, ascorbate-dependent monooxygenases signature 2.

CONSENSUS: H-x-F-x(4)-H-T-H-x(2)-G.

45

NAME: Tyrosinase CuA-binding region signature.

CONSENSUS: H-x(4,5)-F-[LIVMFTP]-x-[FW]-H-R-x(2)-[LM]-x(3)-E.

NAME: Tyrosinase and hemocyanins CuB-binding region signature.

50 CONSENSUS: D-P-x-F-[LIVMFYW]-x(2)-H-x(3)-D.

NAME: Fatty acid desaturases family 1 signature.

CONSENSUS: G-E-x-[FY]-H-N-[FY]-H-H-x-F-P-x-D-Y.

55

NAME: Fatty acid desaturases family 2 signature.

CONSENSUS: [ST]-[SA]-x(3)-[QR]-[LI]-x(5,6)-D-Y-x(2)-[LIVMFYW]-[LIVM]-[DE].

- NAME: Cytochrome P450 cysteine heme-iron ligand signature.
 CONSENSUS: [FW]-[SGNH]-x-[GD]-x-[RHPT]-x-C-[LIVMFAP]-[GAD].
- 5 NAME: Heme oxygenase signature.
 CONSENSUS: L-L-V-A-H-A-Y-T-R.
- NAME: Copper/Zinc superoxide dismutase signature 1.
 CONSENSUS: [GA]-[IFAT]-H-[LIVF]-H-x(2)-[GP]-[SDG]-x-[STAGD].
- 10 NAME: Copper/Zinc superoxide dismutase signature 2.
 CONSENSUS: G-[GN]-[SGA]-G-x-R-x-[SGA]-C-x(2)-[IV].
- NAME: Manganese and iron superoxide dismutases signature.
 CONSENSUS: D-x-W-E-H-[STA]-[FY](2).
- 15 NAME: Ribonucleotide reductase large subunit signature.
 CONSENSUS: W-x(2)-[LF]-x(6,7)-G-[LIVM]-[FYRA]-[NH]-x(3)-
 [STAQLIVM]-[ASC]-x(2)-
 CONSENSUS: [PA].
- 20 NAME: Ribonucleotide reductase small subunit signature.
 CONSENSUS: [IVMSEQ]-E-x(1,2)-[LIVTA]-[HY]-[GSA]-x-[STAVM]-Y-
 x(2)-[LIVMQ]-x(3)-
 CONSENSUS: [LIFY]-[IVFYCSA].
- 25 NAME: Nitrogenases component 1 alpha and beta subunits
 signature 1.
 CONSENSUS: [LIVMFYH]-[LIVMFST]-H-[AG]-[AGSP]-[LIVMNQA]-[AG]-
 C.
- 30 NAME: Nitrogenases component 1 alpha and beta subunits
 signature 2.
 CONSENSUS: [STANQ]-[ET]-C-x(5)-G-D-[DN]-[LIVMT]-x-[STAGR]-
 [LIVMFYST].
- 35 NAME: NifH/frxC family signature 1.
 CONSENSUS: E-x-G-G-P-x(2)-[GA]-x-G-C-[AG]-G.
- NAME: NifH/frxC family signature 2.
 CONSENSUS: D-x-L-G-D-V-V-C-G-G-F-[AG]-x-P.
- 40 NAME: Nickel-dependent hydrogenases large subunit signature
 1.
 CONSENSUS: R-G-[LIVMF]-E-x(15)-[QESM]-R-x-C-G-[LIVM]-C.
- 45 NAME: Nickel-dependent hydrogenases large subunit signature
 2.
 CONSENSUS: [FY]-D-P-C-[LIM]-[ASG]-C-x(2,3)-H.
- 50 NAME: Glutamyl-tRNA reductase signature.
 CONSENSUS: H-[LIVM]-x(2)-[LIVM]-[GSTAC](3)-[LIVM]-[DEQ]-S-
 [LIVMA]-[LIVM](2)-[GF]-E-
 CONSENSUS: x-[QR]-[IV]-[LIT]-[STAG]-Q-[LIVM]-[EKR].
- 55 NAME: Bacterial-type phytoene dehydrogenase signature.
 CONSENSUS: [NG]-x-[FYVV]-[LIVMF]-x-G-[AGC]-[GS]-[TA]-[HQT]-P-
 G-[STAV]-G-[LIVM]-
 CONSENSUS: x(5)-[GS].

- NAME: Glycine radical signature.
 CONSENSUS: [ESTIV]-x-R-[IVT]-[CSA]-G-Y-x-[GACV].
- 5 NAME: Ergosterol biosynthesis ERG4/ERG24 family signature 1.
 CONSENSUS: G-x(2)-[LIVM]-Y-D-x-[FY]-x-G-x(2)-L-N-P-R.
- NAME: Ergosterol biosynthesis ERG4/ERG24 family signature 2.
 CONSENSUS: [LIVM](2)-H-R-x(2)-R-D-x(3)-C-x(2)-K-Y-G.
- 10 NAME: NNMT/PNMT/TEMT family of methyltransferases signature.
 CONSENSUS: L-I-D-I-G-S-G-P-T-[IV]-Y-Q-L-L-S-A-C.
- NAME: RNA methyltransferase trmA family signature 1.
 15 CONSENSUS: [DN]-P-[PA]-R-x-G-x(14,16)-[LIVM](2)-Y-x-S-C-N-x(2)-T.
- NAME: RNA methyltransferase trmA family signature 2.
 CONSENSUS: [LIVMF]-D-x-F-P-[QHY]-[ST]-x-H-[LIVMFY]-E.
- 20 NAME: Thymidylate synthase active site.
 CONSENSUS: R-x(2)-[LIVM]-x(3)-[FW]-[QN]-x(8,9)-[LV]-x-P-C-[HAVM]-x(3)-[QMT]-[FYW]-
 CONSENSUS: x-[LV].
- 25 NAME: Ribosomal RNA adenine dimethylases signature.
 CONSENSUS: [LIVM]-[LIVMFY]-[DE]-x-G-[STAPV]-G-x-[GA]-x-[LIVMF]-[ST]-x(2)-[LIVM]-
 CONSENSUS: x(6)-[LIVMY]-x-[STAGV]-[LIVMFYHC]-E-x-D.
- 30 NAME: Methylated-DNA--protein-cysteine methyltransferase active site.
 CONSENSUS: [LIVMF]-P-C-H-R-[LIVMF](2).
- 35 NAME: N-6 Adenine-specific DNA methylases signature.
 CONSENSUS: [LIVMAC]-[LIVFYWA]-x-[DN]-P-P-[FYW].
- NAME: N-4 cytosine-specific DNA methylases signature.
 CONSENSUS: [LIVMF]-T-S-P-P-[FY].
- 40 NAME: C-5 cytosine-specific DNA methylases active site.
 CONSENSUS: [DENKS]-x-[FLIV]-x(2)-[GSTC]-x-P-C-x(2)-[FYWLIM]-S.
- 45 NAME: C-5 cytosine-specific DNA methylases C-terminal signature.
 CONSENSUS: [RKQGTF]-x(2)-G-N-[STAG]-[LIVMF]-x(3)-[LIVMT]-x(3)-[LIVM]-x(3)-[LIVM].
- 50 NAME: Protein-L-isoaspartate(D-aspartate) O-methyltransferase signature.
 CONSENSUS: [GSA]-D-G-x(2)-G-[FYWV]-x(3)-[AS]-P-[FY]-[DN]-x-I.
- NAME: Uroporphyrin-III C-methyltransferase signature 1.
 55 CONSENSUS: [LIVM]-[GS]-[STAL]-G-P-G-x(3)-[LIVMFY]-[LIVM]-T-[LIVM]-[KRRHQG]-[AG].
- NAME: Uroporphyrin-III C-methyltransferase signature 2.

CONSENSUS: V-x(2)-[LI]-x(2)-G-D-x(3)-[FYW]-[GS]-x(8)-[LIVF]-
x(5,6)-[LIVMFYWPAC]-
CONSENSUS: x-[LIVMY]-x-P-G.

5 NAME: ubiE/C0Q5 methyltransferase family signature 1.
CONSENSUS: Y-D-x-M-N-x(2)-[LIVM]-S-x(3)-H-x(2)-W.

NAME: ubiE/C0Q5 methyltransferase family signature 2.
CONSENSUS: R-V-[LIVM]-K-[PV]-G-G-x-[LIVMF]-x(2)-[LIVM]-E-x-S.

10 NAME: Serine hydroxymethyltransferase pyridoxal-phosphate
attachment site.
CONSENSUS: [DEH]-[LIVMFY]-x-[STMV]-[GST]-[ST](2)-H-K-[ST]-
[LF]-x-G-[PAC]-[RQ]-
15 CONSENSUS: [GSA]-[GA].

NAME: Phosphoribosylglycinamide formyltransferase active
site.
CONSENSUS: G-x-[STM]-[IVT]-x-[FYWVQ]-[VMAT]-x-[DEV]-x-
20 [LIVMY]-D-x-G-x(2)-[LIVT]-
CONSENSUS: x(6)-[LIVM].

NAME: Aspartate and ornithine carbamoyltransferases
signature.
25 CONSENSUS: F-x-[EK]-x-S-[GT]-R-T.

NAME: Transketolase signature 1.
CONSENSUS: R-x(3)-[LIVMTA]-[DENQSTHKF]-x(5,6)-[GSN]-G-H-
[PLIVMF]-[GSTA]-x(2)-
30 CONSENSUS: [LIMC]-[GS].

NAME: Transketolase signature 2.
CONSENSUS: G-[DEQSA]-[DN]-G-[PAEQ]-[ST]-[HQ]-x-[PAGM]-
[LIVMYAC]-[DEFYU]-x(2)-
35 CONSENSUS: [STAP]-x(2)-[RGA].

NAME: Transaldolase signature 1.
CONSENSUS: [DG]-[IVSA]-T-[ST]-N-P-[STA]-[LIVMF](2).

40 NAME: Transaldolase active site.
CONSENSUS: [LIVM]-x-[LIVM]-K-[LIVM]-[PAS]-x-[ST]-x-[DENQPAS]-
G-[LIVM]-x-[AGV]-x-
CONSENSUS: [QEKIRST]-x-[LIVM].

45 NAME: Acyltransferases ChoActase / COT / CPT family
signature 1.
CONSENSUS: [LI]-P-x-[LVP]-P-[IVTA]-P-x-[LIVM]-x-[DENQAS]-
[ST]-[LIVM]-x(2)-[LY].

50 NAME: Acyltransferases ChoActase / COT / CPT family
signature 2.
CONSENSUS: R-[FYW]-x-[DA]-[KA]-x(0,1)-[LIVMFY]-x-[LIVMFY](2)-
x(3)-[DNS]-[GSA]-x(6)-
CONSENSUS: [DE]-[HS]-x(3)-[DE]-[GA].

55 NAME: Thiolases acyl-enzyme intermediate signature.
CONSENSUS: [LIVM]-[NST]-x(2)-C-[SAGLI]-[ST]-[SAG]-[LIVMFYNS]-
x-[TAG]-[LIVM]-x(6)-

- CONSENSUS: [LIVM].
- NAME: Thiolases signature 2.
 5 CONSENSUS: N-x(2)-G-G-x-[LIVM]-[SA]-x-G-H-P-x-G-x-[ST]-G.
- NAME: Thiolases active site.
 CONSENSUS: [AG]-[LIVMA]-[STAGLIVM]-[STAG]-[LIVMA]-C-x-[AG]-x-[AG]-x-[AG]-x-[SAG].
- 10 NAME: Chloramphenicol acetyltransferase active site.
 CONSENSUS: Q-[LIV]-H-H-[SA]-x(2)-D-G-[FY]-H.
- NAME: Hexapeptide-repeat containing-transferases signature.
 15 CONSENSUS: [LIV]-[GAED]-x(2)-[STAV]-x-[LIV]-x(3)-[LIVAC]-x-[LIV]-[GAED]-x(2)-
 CONSENSUS: [STAVR]-x-[LIV]-[GAED]-x(2)-[STAV]-x-[LIV]-x(3)-[LIV].
- NAME: Beta-ketoacyl synthases active site.
 20 CONSENSUS: G-x(4)-[LIVMFAP]-x(2)-[AGC]-C-[STA](2)-[STAG]-x(3)-[LIVMF].
- NAME: Chalcone and stilbene synthases active site.
 25 CONSENSUS: R-[LIVMFYS]-x-[LIVM]-x-[QHG]-x-G-C-[FYNA]-[GA]-G-[GA]-[STAV]-x-[LIVMF]-
 CONSENSUS: [RA].
- NAME: Myristoyl-CoA:protein N-myristoyltransferase signature 1.
 30 CONSENSUS: E-I-N-F-L-C-x-H-K.
- NAME: Myristoyl-CoA:protein N-myristoyltransferase signature 2.
 35 CONSENSUS: K-F-G-x-G-D-G.
- NAME: Gamma-glutamyltranspeptidase signature.
 CONSENSUS: T-[STA]-H-x-[ST]-[LIVMA]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-[LIVM]-[NE]-
 40 CONSENSUS: x(1,2)-[FY]-G.
- NAME: Transglutaminases active site.
 CONSENSUS: [GT]-Q-[CA]-W-V-x-[SA]-[GA]-[IVT]-x(2)-T-x-[LMSC]-R-[CSA]-[LV]-G.
- NAME: Phosphorylase pyridoxal-phosphate attachment site.
 45 CONSENSUS: E-A-[SC]-G-x-[GS]-x-M-K-x(2)-[LM]-N.
- NAME: UDP-glycosyltransferases signature.
 50 CONSENSUS: [FW]-x(2)-Q-x(2)-[LIVMYA]-[LIMV]-x(4,6)-[LVGAC]-[LVFYA]-[LIVMF]-[STAGCM]-
 CONSENSUS: [HNQ]-[STAGC]-G-x(2)-[STAG]-x(3)-[STAGL]-[LIVMFA]-x(4)-[PQR]-[LIVMT]-
 CONSENSUS: x(3)-[PA]-x(3)-[DES]-[QEHN].
- NAME: Purine/pyrimidine phosphoribosyl transferases signature.
 55 CONSENSUS: [LIVMFYWCTA]-[LIVM]-[LIVMA]-[LIVMFC]-[DE]-D-[LIVMS]-[LIVM]-[STAVD]-

CONSENSUS: [ESTAR]-[GAC]-x-[ESTAR].

NAME: Glutamine amidotransferases class-I active site.

5 CONSENSUS: [PAS]-[LIVMFYT]-[LIVMFY]-G-[LIVMFY]-C-[LIVMFYN]-G-x-[QEH]-x-[LIVMFA].

NAME: Glutamine amidotransferases class-II active site.

CONSENSUS: <x(0,11)-C-[GS]-[IV]-[LIVMFYW]-[AG].

10 NAME: Purine and other phosphorylases family 1 signature.

CONSENSUS: [GST]-x-G-[LIVM]-G-x-[PA]-S-x-[GSTA]-I-x(3)-E-L.

NAME: Purine and other phosphorylases family 2 signature.

15 CONSENSUS: [LIV]-x(3)-G-x(2)-H-x-[LIVMFY]-x(4)-[LIVMF]-x(3)-[ATV]-x(1,2)-[LIVM]-x-
CONSENSUS: [ATV]-x(4)-[GN]-x(3,4)-[LIVMF](2)-x(2)-[STN]-[SA]-x-G-[GS]-[LIVM].

20 NAME: Thymidine and pyrimidine-nucleoside phosphorylases signature.

CONSENSUS: S-[GS]-R-[GA]-[LIV]-x(2)-[TA]-[GA]-G-T-x-D-x-[LIV]-E.

NAME: ATP phosphoribosyltransferase signature.

25 CONSENSUS: E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM].

NAME: NAD:arginine ADP-ribosyltransferases signature.

30 CONSENSUS: [FY]-x-[FY]-K-x(2)-H-[FY]-x-L-[ST]-x-A.

NAME: Prolipoprotein diacylglycerol transferase signature.

CONSENSUS: G-R-x-[GA]-N-F-[LIVMF]-N-x-E-x(2)-G.

NAME: S-adenosylmethionine synthetase signature 1.

35 CONSENSUS: G-A-G-D-Q-G-x(3)-G-Y.

NAME: S-adenosylmethionine synthetase signature 2.

CONSENSUS: G-[GA]-G-[ASC]-F-S-x-K-[DE].

40 NAME: Polyprenyl synthetases signature 1.

CONSENSUS: [LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH].

NAME: Polyprenyl synthetases signature 2.

45 CONSENSUS: [LIVMFY]-G-x(2)-[FY]-Q-[LIVM]-x-D-D-[LIVMFY]-x-[DNG].

NAME: Squalene and phytoene synthases signature 1.

50 CONSENSUS: Y-[CSAM]-x(2)-[VSG]-A-[GSA]-[LIVAT]-[IV]-G-x(2)-[LMSC]-x(2)-[LIV].

NAME: Squalene and phytoene synthases signature 2.

55 CONSENSUS: [LIVM]-G-x(3)-Q-x(2,3)-N-[IF]-x-R-D-[LIVMFY]-x(2)-[DE]-x(4,7)-R-x-[FY]-
CONSENSUS: x-P.

NAME: Protein prenyltransferases alpha subunit repeat signature.

CONSENSUS: [PSIAV]-x-[NDFV]-[NEQIY]-x-[LIVMAGP]-W-[NQSTHF]-
[FYHQ]-[LIVMR].

NAME: Riboflavin synthase alpha chain family signature.

5 CONSENSUS: [LIVMF]-x(5)-G-[STADNQ]-[KREQIYW]-V-N-[LIVM]-E.

NAME: Dihydropteroate synthase signature 1.

CONSENSUS: [LIVM]-x-[AG]-[LIVMF](2)-N-x-T-x-D-S-F-x-D-x-[SG].

10 NAME: Dihydropteroate synthase signature 2.

CONSENSUS: [GE]-[SA]-x-[LIVM](2)-D-[LIVM]-G-[GP]-x(2)-[STA]-
x-P.

NAME: EPSP synthase signature 1.

15 CONSENSUS: [LIVM]-x(2)-[GN]-N-[SA]-G-T-[STA]-x-R-x-[LIVMY]-x-
[GSTA].

NAME: EPSP synthase signature 2.

20 CONSENSUS: [KR]-x-[KH]-E-[CST]-[DNE]-R-[LIVM]-x-[STA]-
[LIVMC]-x(2)-[EN]-[LIVMF]-x-
CONSENSUS: [KRA]-[LIVMF]-G.

NAME: FLAP/GST2/LTC4S family signature.

25 CONSENSUS: G-x(3)-F-E-R-V-[FY]-x-A-[NQ]-x-N-C.

NAME: Aminotransferases class-I pyridoxal-phosphate
attachment site.

CONSENSUS: [GS]-[LIVMFYTAC]-[GSTA]-K-x(2)-[GSALVN]-[LIVMFA]-
x-[GNAR]-x-R-[LIVMA]-

30 CONSENSUS: [GA].

NAME: Aminotransferases class-II pyridoxal-phosphate
attachment site.

35 CONSENSUS: T-[LIVMFYW]-[STAG]-K-[SAG]-[LIVMFYWR]-[SAG]-x(2)-
[SAG].

NAME: Aminotransferases class-III pyridoxal-phosphate
attachment site.

40 CONSENSUS: [LIVMFYW](2)-x-D-E-[LIVMA]-x(2)-[GP]-x(0,1)-
[LIVMFYWAG]-x(0,1)-[SACR]-x-
CONSENSUS: [GSAD]-x(12,16)-D-[LIVMFYW](2,3)-[GSA]-K-x(3)-
[GSTADN]-[GSA].

NAME: Aminotransferases class-IV signature.

45 CONSENSUS: E-x-[STAGCI]-x(2)-N-[LIVMFAC]-[FY]-x(6,12)-
[LIVMF]-x-T-x(6,8)-[LIVM]-x-
CONSENSUS: [GS]-[LIVM]-x-[KR].

NAME: Aminotransferases class-V pyridoxal-phosphate
attachment site.

50 CONSENSUS: [LIVFYCHT]-[DGH]-[LIVMFYAC]-[LIVMFYA]-x(2)-
[GSTAC]-[GSTA]-[HQR]-K-
CONSENSUS: x(4,6)-G-x-[GSAT]-x-[LIVMFYSAC].

55 NAME: Hexokinases signature.

CONSENSUS: [LIVM]-G-F-[TN]-F-S-[FY]-P-x(5)-[LIVM]-[DNST]-
x(3)-[LIVM]-x(2)-W-T-K-x-
CONSENSUS: [LF].

- NAME: Galactokinase signature.
 CONSENSUS: G-R-x-N-[LIV]-I-G-E-H-x-D-Y.
- 5 NAME: GHMP kinases putative ATP-binding domain.
 CONSENSUS: [LIVM]-[PK]-x-[GSTA]-x(0,1)-G-L-[GS]-S-S-[GSA]-[GSTAC].
- 10 NAME: Phosphofructokinase signature.
 CONSENSUS: [RK]-x(4)-G-H-x-Q-[QR]-G-G-x(5)-D-R.
- NAME: pfkB family of carbohydrate kinases signature 1.
 CONSENSUS: [AG]-G-x(0,1)-[GAP]-x-N-x-[STA]-x(6)-[GS]-x(9)-G.
- 15 NAME: pfkB family of carbohydrate kinases signature 2.
 CONSENSUS: [DNSK]-[PSTV]-x-[SAG](2)-[GD]-D-x(3)-[SAGV]-[AG]-[LIVMFY]-[LIVMSTAP].
- 20 NAME: ROK family signature.
 CONSENSUS: [LIVM]-x(2)-G-[LIVMFCT]-G-x-[GA]-[LIVMFA]-x(8)-G-x(3,5)-[GATP]-x(2)-
 CONSENSUS: G-[RKH].
- 25 NAME: Phosphoribulokinase signature.
 CONSENSUS: K-[LIVM]-x-R-D-x(3)-R-G-x-[ST]-x-E.
- NAME: Thymidine kinase cellular-type signature.
 CONSENSUS: [GA]-x(1,2)-[DE]-x-Y-x-[STAP]-x-C-[NKR]-x-[CH]-[LIVMFYWH].
- 30 NAME: FGGY family of carbohydrate kinases signature 1.
 CONSENSUS: [MFYGS]-x-[PST]-x(2)-K-[LIVMFYW]-x-W-[LIVMF]-x-[DENQTKR]-[ENQH].
- 35 NAME: FGGY family of carbohydrate kinases signature 2.
 CONSENSUS: [GSA]-x-[LIVMFYW]-x-G-[LIVM]-x(7,8)-[HDENQ]-[LIVMF]-x(2)-[AS]-[STAIVM]-
 CONSENSUS: [LIVMFY]-[DEQ].
- 40 NAME: Protein kinases ATP-binding region signature.
 CONSENSUS: [LIV]-G-[P]-G-[P]-[FYWMGSTNH]-[SGA]-[PW]-[LIVCAT]-[PD]-x-[GSTACLIVMFY]-
 CONSENSUS: x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K.
- 45 NAME: Serine/Threonine protein kinases active-site signature.
 CONSENSUS: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3).
- 50 NAME: Tyrosine protein kinases specific active-site signature.
 CONSENSUS: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTAC]-x(2)-N-[LIVMFYC](3).
- 55 NAME: Protein kinase domain profile.
 NAME: Casein kinase II regulatory subunit signature.

CONSENSUS: C-P-x-[LIVMY]-x-C-x(5)-L-P-[LIVMC]-G-x(9)-V-[KR]-
x(2)-C-P-x-C.

NAME: Pyruvate kinase active site signature.

5 CONSENSUS: [LIVAC]-x-[LIVM](2)-[SAPCV]-K-[LIV]-E-[NKRST]-x-
[DEQH]-[GSTA]-[LIVM].

NAME: Shikimate kinase signature.

10 CONSENSUS: [KR]-x(2)-E-x(3)-[LIVMF]-x(8,12)-[LIVMF](2)-[SA]-
x-G(3)-x-[LIVMF].

NAME: Prokaryotic diacylglycerol kinase signature.

CONSENSUS: E-x-[LIVM]-N-[EST]-[SA]-[LIV]-E-x(2)-V-D.

15 NAME: Phosphatidylinositol 3- and 4-kinases signature 1.

CONSENSUS: [LIVMFAC]-K-x(1,3)-[DEA]-[DE]-[LIVMC]-R-Q-[DE]-
x(4)-Q.

NAME: Phosphatidylinositol 3- and 4-kinases signature 2.

20 CONSENSUS: [GS]-x-[AV]-x(3)-[LIVM]-x(2)-[FYH]-[LIVM](2)-x-
[LIVMF]-x-D-R-H-x(2)-N.

NAME: Acetate and butyrate kinases family signature 1.

25 CONSENSUS: [LIVM](2)-x-[LIVM]-N-x-G-S-[EST]-S-x-[KE].

NAME: Acetate and butyrate kinases family signature 2.

CONSENSUS: [LIVMA](2)-x(2)-H-x-G-x-G-x-[EST]-[LIVM]-x-[AV]-
x(3)-G.

30 NAME: Phosphoglycerate kinase signature.

CONSENSUS: [KRHGTCV]-[VT]-[LIVMF]-[LIVMC]-R-x-D-x-N-[SACV]-P.

NAME: Aspartokinase signature.

35 CONSENSUS: [LIVM]-x-K-[FY]-G-G-[EST]-[SC]-[LIVM].

NAME: Glutamate 5-kinase signature.

CONSENSUS: [GSTN]-x(2)-G-x-G-[GC]-[IM]-x-[STA]-K-[LIVM]-x-
[SA]-[TCA]-x(2)-[GALV]-
40 CONSENSUS: x(3)-G.

NAME: ATP:guano phosphotransferases active site.

CONSENSUS: C-P-x(0,1)-[EST]-N-[IL]-G-T.

45 NAME: PTS HPR component histidine phosphorylation site
signature.

CONSENSUS: G-[LIVM]-H-[STA]-R-[PA]-[GSTA]-[ESTAM].

NAME: PTS HPR component serine phosphorylation site
signature.

50 CONSENSUS: [GSADE]-[KREQTV]-x(4)-[KRN]-S-[LIVMF](2)-x-[LIVM]-
x(2)-[LIVM]-[GAD].

NAME: PTS EIIA domains phosphorylation site signature 1.

55 CONSENSUS: G-x(2)-[LIVMF](3)-H-[LIVMF]-G-[LIVMF]-x-T-[ALV].

NAME: PTS EIIA domains phosphorylation site signature 2.

CONSENSUS: [DENQ]-x(6)-[LIVMF]-[GA]-x(2)-[LIVM]-A-[LIVM]-P-H-
[GAC].

- NAME: PTS EIIB domains cysteine phosphorylation site
signature.
5 CONSENSUS: N-[LIVMFY]-x(5)-C-x-T-R-[LIVMF]-x-[LIVMF]-x-[LIVM]-x-[DQ].
- NAME: Adenylate kinase signature.
CONSENSUS: [LIVMFYW](3)-D-G-[FYI]-P-R-x(3)-[NQ].
- 10 NAME: Nucleoside diphosphate kinases active site.
CONSENSUS: N-x(2)-H-[GA]-S-D-[SA]-[LIVMPKNE].
- NAME: Guanylate kinase signature.
15 CONSENSUS: T-[ST]-R-x(2)-[KR]-x(2)-[DE]-x(2)-G-x(2)-Y-x-[FY]-[LIVMK].
- NAME: Guanylate kinase domain profile.
- 20 NAME: Phosphoribosyl pyrophosphate synthetase signature.
CONSENSUS: D-[LI]-H-[SA]-x-Q-[IMST]-[QM]-G-[FY]-F-x(2)-P-[LIVMFC]-D.
- NAME: 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase
signature.
25 CONSENSUS: G-[PE]-R-x(2)-D-L-D-[LIVM](2).
- NAME: Bacteriophage-type RNA polymerase family active site
signature 1.
30 CONSENSUS: P-[LIVM]-x(2)-D-[GA]-[ST]-[AC]-[SN]-[GA]-[LIVMFY]-Q.
- NAME: Bacteriophage-type RNA polymerase family active site
signature 2.
35 CONSENSUS: [LIVMF]-x-R-x(3)-K-x(2)-[LIVMF]-M-[PT]-x(2)-Y.
- NAME: Eukaryotic RNA polymerase II heptapeptide repeat.
CONSENSUS: Y-[ST]-P-[ST]-S-P-[STANK].
- 40 NAME: RNA polymerases beta chain signature.
CONSENSUS: G-x-K-[LIVMFA]-[STAC]-[GSTN]-x-[HSTA]-[GS]-[QNH]-K-G-[IVT].
- NAME: RNA polymerases M / 15 Kd subunits signature.
45 CONSENSUS: F-C-x-[DEKST]-C-[GNK]-[DNSA]-[LIVMH]-[LIVM]-x(8,14)-C-x(2)-C.
- NAME: RNA polymerases D / 30 to 40 Kd subunits signature.
CONSENSUS: N-[SGA]-[LIVMF]-R-R-x(9)-[SA]-x(3)-V-x(4)-N-x-[STA]-x(3)-[DN]-E-x-[LI]-
50 CONSENSUS: [GA]-x-R-[LI]-[GA]-[LIVM](2)-P.
- NAME: RNA polymerases H / 23 Kd subunits signature.
CONSENSUS: H-[NEI]-[LIVM]-V-P-x-H-x(2)-[LIVM]-x(2)-[DE].
- 55 NAME: RNA polymerases K / 14 to 18 Kd subunits signature.
CONSENSUS: [ST]-x-[FY]-E-x-[AT]-R-x-[LIVM]-[GSA]-x-R-[SA]-x-Q.

- NAME: RNA polymerases L / 13 to 16 Kd subunits signature.
 CONSENSUS: [DEJ](2)-H-[EST]-[LIVM]-[GAP]-N-x(11)-V-x-[FM]-x(2)-Y-x(3)-H-P.
- 5 NAME: RNA polymerases N / 8 Kd subunits signature.
 CONSENSUS: [LIVMF](2)-P-[LIVM]-x-C-F-[EST]-C-G.
- NAME: DNA polymerase family A signature.
 CONSENSUS: R-x(2)-[GSAV]-K-x(3)-[LIVMFY]-[AGQ]-x(2)-Y-x(2)-[GS]-x(3)-[LIVMA].
- 10 NAME: DNA polymerase family B signature.
 CONSENSUS: [YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-x-[LIVMSTAC].
- 15 NAME: DNA polymerase family X signature.
 CONSENSUS: G-[SG]-[LFY]-x-R-[GE]-x(3)-[SGCL]-x-D-[LIVM]-D-[LIVMFY](3)-x(2)-[SAP].
- 20 NAME: Galactose-1-phosphate uridyl transferase family 1 active site signature.
 CONSENSUS: F-E-N-[RK]-G-x(3)-G-x(4)-H-P-H-x-Q.
- NAME: Galactose-1-phosphate uridyl transferase family 2 signature.
 25 CONSENSUS: D-L-P-I-V-G-G-[EST]-[LIVM](2)-[SA]-H-[DEN]-H-[FY]-Q-G-G.
- NAME: ADP-glucose pyrophosphorylase signature 1.
 30 CONSENSUS: [AG]-G-G-x-G-[STK]-x-L-x(2)-L-[TA]-x(3)-A-x-P-A-[LV].
- NAME: ADP-glucose pyrophosphorylase signature 2.
 CONSENSUS: W-[FY]-x-G-[EST]-A-[DNSH]-[AS]-[LIVMFYW].
- 35 NAME: ADP-glucose pyrophosphorylase signature 3.
 CONSENSUS: [APV]-[GS]-M-G-[LIVMN]-Y-[IVC]-[LIVMFY]-x(2)-[DENPHK].
- NAME: Phosphatidate cytidyltransferase signature.
 40 CONSENSUS: S-x-[LIVMF]-K-R-x(4)-K-D-x-[GSA]-x(2)-[LI]-[PG]-x-H-G-G-[LIVM]-x-D-R-
 CONSENSUS: [LIVMFT]-D.
- NAME: Ribonuclease PH signature.
 45 CONSENSUS: C-[DEJ]-[LIVM](2)-Q-[GTA]-D-G-[SG]-x(2)-[TA]-A.
- NAME: 2'-5'-oligoadenylate synthetases signature 1.
 CONSENSUS: G-G-S-x-[AG]-[KR]-x-T-x-L-[KR]-[GST]-x-S-D-[AG].
- 50 NAME: 2'-5'-oligoadenylate synthetases signature 2.
 CONSENSUS: R-P-V-I-L-D-P-x-[DEJ]-P-T.
- NAME: CDP-alcohol phosphatidyltransferases signature.
 55 CONSENSUS: D-G-x(2)-A-R-x(8)-G-x(3)-D-x(3)-D.
- NAME: PEP-utilizing enzymes phosphorylation site signature.

- CONSENSUS: G-[GA]-x-[TN]-x-H-[STA]-[STAV]-[LIVM](2)-[STAV]-[RG].
 NAME: PEP-utilizing enzymes signature 2.
 5 CONSENSUS: [DEQS]-x-[LIVMF]-S-[LIVMF]-G-[ST]-N-D-[LIVM]-x-Q-[LIVMFYGT]-[STALIV]-
 CONSENSUS: [LIVMF]-[GAS]-x(2)-R.
- 10 NAME: Rhodanese signature 1.
 CONSENSUS: [FY]-x(3)-H-[LIV]-P-G-A-x(2)-[LIVF].
 NAME: Rhodanese C-terminal signature.
 CONSENSUS: [AV]-x(2)-[FY]-[DEAP]-G-[GSA]-[WF]-x-E-[FYW].
- 15 NAME: CoA transferases signature 1.
 CONSENSUS: [DN]-[GN]-x(2)-[LIVMFA](3)-G-G-F-x(3)-G-x-P.
 NAME: CoA transferases signature 2.
 20 CONSENSUS: [LF]-[HQ]-S-E-N-G-[LIVF](2)-[GA].
 NAME: Phospholipase A2 histidine active site.
 CONSENSUS: C-C-x(2)-H-x(2)-C.
- 25 NAME: Phospholipase A2 aspartic acid active site.
 CONSENSUS: [LIVMA]-C-[LIVMFYWPCT]-C-D-x(5)-C.
 NAME: Lipases, serine active site.
 CONSENSUS: [LIV]-x-[LIVFY]-[LIVMST]-G-[HYWV]-S-x-G-[GSTAC].
- 30 NAME: Colipase signature.
 CONSENSUS: Y-x(2)-Y-Y-x-C-x-C.
 NAME: Lipolytic enzymes "G-D-S-L" family, serine active site.
 35 CONSENSUS: [LIVMFYAG](4)-G-D-S-[LIVM]-x(1,2)-[TAG]-G.
 NAME: Lipolytic enzymes "G-D-X-G" family, putative histidine active site.
 CONSENSUS: [LIVMF](2)-x-[LIVMF]-H-G-G-[SAG]-[FY]-x(3)-[STDN]-x(2)-[ST]-H.
 40 NAME: Lipolytic enzymes "G-D-X-G" family, putative serine active site.
 CONSENSUS: [LIVM]-x-[LIVMF]-[SA]-G-D-S-[CA]-G-[GA]-x-L-[CA].
- 45 NAME: Carboxylesterases type-B serine active site.
 CONSENSUS: F-[GR]-G-x(4)-[LIVM]-x-[LIV]-x-G-x-S-[STAG]-G.
 NAME: Carboxylesterases type-B signature 2.
 50 CONSENSUS: [ED]-D-C-L-[YT]-[LIV]-[DNS]-[LIV]-[LIVFYW]-x-[PQR].
 NAME: Pectinesterase signature 1.
 CONSENSUS: [GSTN]-x(5)-[LIVM]-x-[LIVM]-x(2)-G-x-Y-[DNK]-E-x-[LIVM]-x-[LIVM].
 55 NAME: Pectinesterase signature 2.
 CONSENSUS: G-[STAD]-[LIVMT]-D-F-I-F-G.

- NAME: Peptidyl-tRNA hydrolase signature 1.
 CONSENSUS: [FY]-x(2)-T-R-H-N-x-G-x(2)-[LIVMFA](2)-[DE].
- 5 NAME: Peptidyl-tRNA hydrolase signature 2.
 CONSENSUS: [GS]-x(3)-H-N-G-[LIVM]-[KR]-[DNS]-[LIVMT].
- NAME: Alkaline phosphatase active site.
 CONSENSUS: [IV]-x-D-S-[GAS]-[GASC]-[GAST]-[GA]-T.
- 10 NAME: Histidine acid phosphatases phosphohistidine
 signature.
 CONSENSUS: [LIVM]-x(2)-[LIVMA]-x(2)-[LIVM]-x-R-H-[GN]-x-R-x-
 [PAS].
- 15 NAME: Histidine acid phosphatases active site signature.
 CONSENSUS: [LIVMF]-x-[LIVMFAG]-x(2)-[STAGI]-H-D-[STANQ]-x-
 [LIVM]-x(2)-[LIVMFY]-x(2)-
 CONSENSUS: [STA].
- 20 NAME: Class A bacterial acid phosphatases signature.
 CONSENSUS: G-S-Y-P-S-G-H-T.
- NAME: 5'-nucleotidase signature 1.
 25 CONSENSUS: [LIVM]-x-[LIVM](2)-[HEA]-[TI]-x-D-x-H-[GSA]-x-
 [LIVMF].
- NAME: 5'-nucleotidase signature 2.
 CONSENSUS: [FYP]-x(4)-[LIVM]-G-N-H-E-F-[DN].
- 30 NAME: Fructose-1,6-bisphosphatase active site.
 CONSENSUS: [AG]-[RK]-L-x(1,2)-[LIV]-[FY]-E-x(2)-P-[LIVM]-
 [GSA].
- 35 NAME: Serine/threonine specific protein phosphatases
 signature.
 CONSENSUS: [LIVM]-R-G-N-H-E.
- NAME: Protein phosphatase 2A regulatory subunit PR55
 signature 1.
 40 CONSENSUS: E-F-D-Y-L-K-S-L-E-I-E-E-K-I-N.
- NAME: Protein phosphatase 2A regulatory subunit PR55
 signature 2.
 45 CONSENSUS: N-[AG]-H-[TA]-Y-H-I-N-S-I-S-[LIVM]-N-S-D.
- NAME: Protein phosphatase 2C signature.
 CONSENSUS: [LIVMFY]-[LIVMFYA]-[GSAC]-[LIVM]-[FYC]-D-G-H-
 [GAV].
- 50 NAME: Tyrosine specific protein phosphatases active site.
 CONSENSUS: [LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY].
- NAME: Tyrosine specific protein phosphatases profile.
- 55 NAME: Dual specificity protein phosphatase profile.
- NAME: PTP type protein phosphatase profile.

- NAME: Inositol monophosphatase family signature 1.
 5 CONSENSUS: [FWV]-x(0,1)-[LIVM]-D-P-[LIVM]-D-[SG]-[ST]-x(2)-[FY]-x-[HKRNSTY].
- NAME: Inositol monophosphatase family signature 2.
 CONSENSUS: [WV]-D-x-[AC]-[GSA]-[GSAPV]-x-[LIVACP]-[LIV]-[LIVAC]-x(3)-[GH]-[GA].
- 10 NAME: Prokaryotic zinc-dependent phospholipase C signature.
 CONSENSUS: H-Y-x-[GT]-D-[LIVM]-[DNS]-x-P-x-H-[PA]-x-N.
- NAME: Phosphatidylinositol-specific phospholipase X-box domain profile.
- 15 NAME: Phosphatidylinositol-specific phospholipase Y-box domain profile.
- NAME: 3'5'-cyclic nucleotide phosphodiesterases signature.
 20 CONSENSUS: H-D-[LIVMFY]-x-H-x-[AG]-x(2)-[NQ]-x-[LIVMFY].
- NAME: cAMP phosphodiesterases class-II signature.
 CONSENSUS: H-x-H-L-D-H-[LIVM]-x-[GS]-[LIVMA]-[LIVM](2)-x-S-[AP].
- 25 NAME: Sulfatases signature 1.
 CONSENSUS: [SAP]-[LIVMST]-[CS]-[STAC]-P-[STA]-R-x(2)-[LIVFW](2)-[TR]-G.
- 30 NAME: Sulfatases signature 2.
 CONSENSUS: G-[YV]-x-[ST]-x(2)-[IVA]-G-K-x(0,1)-[FYWK]-[HL].
- NAME: AP endonucleases family 1 signature 1.
 CONSENSUS: [APF]-D-[LIVMF](2)-x-[LIVM]-Q-E-x-K.
- 35 NAME: AP endonucleases family 1 signature 2.
 CONSENSUS: D-[ST]-[FY]-R-[KH]-x(7,8)-[FYW]-[ST]-[FYW](2).
- NAME: AP endonucleases family 1 signature 3.
 40 CONSENSUS: N-x-G-x-R-[LIVM]-D-[LIVMFYH]-x-[LV]-x-S.
- NAME: AP endonucleases family 2 signature 1.
 CONSENSUS: H-x(2)-Y-[LIVMF]-[IM]-N-[LIVMCA]-[AG].
- 45 NAME: AP endonucleases family 2 signature 2.
 CONSENSUS: [GR]-[LIVMF]-C-[LIVM]-D-T-C-H.
- NAME: AP endonucleases family 2 signature 3.
 CONSENSUS: [LIVMW]-H-x-N-[DE]-[SA]-K-x(3)-G-[SA]-x(2)-D.
- 50 NAME: Deoxyribonuclease I signature 1.
 CONSENSUS: [LIVM](2)-[AP]-L-H-[STA](2)-P-x(5)-E-[LIVM]-[DN]-x-L-x-[DE]-V.
- 55 NAME: Deoxyribonuclease I signature 2.
 CONSENSUS: G-D-F-N-A-x-C-[SA].
- NAME: Endonuclease III iron-sulfur binding region signature.

- 5
 10
 15
 20
 25
 30
 35
 40
 45
 50
 55
- CONSENSUS: C-x(3)-[KRS]-P-[KRAL]-C-x(2)-C-x(5)-C.
 NAME: Endonuclease III family signature.
 CONSENSUS: [GST]-x-[LIVMF]-P-x(5)-[LIVMW]-x(2,3)-[LI]-[PAS]-
 G-V-[GA]-x(3)-[GAC]-
 CONSENSUS: x(3)-[LIVM]-x(2)-[SALV]-[LIVMFYW]-[GANK].
 NAME: Ribonuclease II family signature.
 CONSENSUS: [HI]-[FYE]-[GSTAM]-[LIVM]-x(4,5)-Y-[STAL]-x-
 [FWVAC]-[TV]-[SA]-P-[LIVMA]-
 CONSENSUS: [RQ]-[KR]-[FY]-x-D-x(3)-[HQ].
 NAME: Ribonuclease III family signature.
 CONSENSUS: [DEQ]-[RQ]-[LM]-E-[FYW]-[LV]-G-D-[SAR].
 NAME: Bacterial Ribonuclease P protein component signature.
 CONSENSUS: [LIVMFYS]-x(2)-A-x(2)-R-[NH]-[KRQL]-[LIVM]-[KRA]-
 R-x-[LIVMTA]-[KR].
 NAME: Ribonuclease T2 family histidine active site 1.
 CONSENSUS: [FYWL]-x-[LIVM]-H-G-L-W-P.
 NAME: Ribonuclease T2 family histidine active site 2.
 CONSENSUS: [LIVMF]-x(2)-[HDGTY]-[EQ]-[FYW]-x-[KR]-H-G-x-C.
 NAME: Pancreatic ribonuclease family signature.
 CONSENSUS: C-K-x(2)-N-T-F.
 NAME: DNA/RNA non-specific endonucleases active site.
 CONSENSUS: D-R-G-H-[QIL]-x(3)-A.
 NAME: Thermonuclease family signature 1.
 CONSENSUS: D-G-D-T-[LIVM]-x-[LIVMC]-x(9,10)-R-[LIVM]-x(2)-
 [LIVM]-D-x-P-E.
 NAME: Thermonuclease family signature 2.
 CONSENSUS: D-[KR]-Y-[GQ]-R-x-[LV]-[GA]-x-[IV]-[FYW].
 NAME: Beta-amylase active site 1.
 CONSENSUS: H-x-C-G-G-N-V-G-D.
 NAME: Beta-amylase active site 2.
 CONSENSUS: G-x-[SA]-G-E-[LIVM]-R-Y-P-S-Y.
 NAME: Glucoamylase active site region signature.
 CONSENSUS: [STN]-[GP]-x(1,2)-[DE]-x-W-E-E-x(2)-[GS].
 NAME: Polygalacturonase active site.
 CONSENSUS: [GSDENKRH]-x(2)-[VMFC]-x(2)-[GS]-H-G-[LIVMAG]-
 x(1,2)-[LIVM]-G-S.
 NAME: Clostridium cellulosome enzymes repeated domain
 signature.
 CONSENSUS: D-[LIVMFY]-[DNV]-x-[DNS]-x(2)-[LIVM]-[DN]-[SALM]-
 x-D-x(3)-[LIVMF]-x-
 CONSENSUS: [RKS]-x-[LIVMF].
 NAME: Chitinases family 1b active site.

CONSENSUS: [LIVMFY]-[DN]-G-[LIVMF]-[DN]-[LIVMF]-[DN]-x-E.

NAME: Chitinases family 19 signature 1.

CONSENSUS: C-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF]-x-A-x(3)-[YF]-
5 x(2)-F-[GSA].

NAME: Chitinases family 19 signature 2.

CONSENSUS: [LIVM]-[GSA]-F-x-[STAG](2)-[LIVMFY]-W-[FY]-W-
10 [LIVM].

NAME: Alpha-lactalbumin / lysozyme C signature.

CONSENSUS: C-x(3)-C-x(2)-[LMF]-x(3)-[DEN]-[LI]-x(5)-C.

NAME: Alpha-galactosidase signature.

CONSENSUS: G-[LIVMFY]-x(2)-[LIVMFY]-x-[LIVM]-D-D-x-W-x(3,4)-
15 R-[DNSF].

NAME: Trehalase signature 1.

CONSENSUS: P-G-G-R-F-x-E-x-Y-x-W-D-x-Y.

NAME: Trehalase signature 2.

CONSENSUS: Q-W-D-x-P-x-[GA]-W-[PA]-P.

NAME: Alpha-L-fucosidase putative active site.

CONSENSUS: P-x(2)-L-x(3)-K-W-E-x-C.

NAME: Glycosyl hydrolases family 1 active site.

CONSENSUS: [LIVMFSTC]-[LIVFYSS]-[LIV]-[LIVMST]-E-N-G-
30 [LIVMFAR]-[CSAGN].

NAME: Glycosyl hydrolases family 1 N-terminal signature.

CONSENSUS: F-x-[FYWM]-[GSTA]-x-[GSTA]-x-[GSTA](2)-[FYNH]-
[NQ]-x-E-x-[GSTA].

NAME: Glycosyl hydrolases family 2 signature 1.

CONSENSUS: N-x-[LIVMFYW]-R-[STACN](2)-H-Y-P-x(4)-

[LIVMFYW](2)-x(3)-[DN]-x(2)-

CONSENSUS: G-[LIVMFYW](4).

NAME: Glycosyl hydrolases family 2 acid/base catalyst.

CONSENSUS: [DENQF]-[KRVW]-N-H-[AP]-[SAC]-[LIVMF](3)-W-[GS]-
40 x(2,3)-N-E.

NAME: Glycosyl hydrolases family 3 active site.

CONSENSUS: [LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-

[LIVMF]-[ST]-D-x(2)-

CONSENSUS: [SGADNI].

NAME: Glycosyl hydrolases family 5 signature.

CONSENSUS: [LIV]-[LIVMFYWGA](2)-[DNEQG]-[LIVMGST]-x-N-E-[PV]-
50 [RHDNSTLIVFY].

NAME: Glycosyl hydrolases family 6 signature 1.

CONSENSUS: V-x-Y-x(2)-P-x-R-D-C-[GSAF]-x(2)-[GSA](2)-x-G.

NAME: Glycosyl hydrolases family 6 signature 2.

CONSENSUS: [LIVMYA]-[LIVA]-[LIVT]-[LIV]-E-P-D-[SAL]-[LI]-
55 [PSAG].

- NAME: Glycosyl hydrolases family 8 signature.
 5 CONSENSUS: A-[EST]-D-[EAG]-D-x(2)-[IM]-A-x-[SA]-[LIVM]-[LIVMG]-
 x-A-x(3)-[FW].
- NAME: Glycosyl hydrolases family 9 active sites signature 1.
 CONSENSUS: [STV]-x-[LIVMFY]-[STV]-x(2)-G-x-[NKR]-x(4)-
 [PLIVM]-H-x-R.
- 10 NAME: Glycosyl hydrolases family 9 active sites signature 2.
 CONSENSUS: [FYW]-x-D-x(4)-[FYW]-x(3)-E-x-[STA]-x(3)-N-[STA].
- NAME: Glycosyl hydrolases family 10 active site.
 15 CONSENSUS: [GTA]-x(2)-[LIVN]-x-[IVMF]-[ST]-E-[LIY]-[DN]-
 [LIVMF].
- NAME: Glycosyl hydrolases family 11 active site signature 1.
 CONSENSUS: [PSA]-[LQ]-x-E-Y-Y-[LIVM](2)-[DE]-x-[FYWHN].
- 20 NAME: Glycosyl hydrolases family 11 active site signature 2.
 CONSENSUS: [LIVMF]-x(2)-E-[AG]-[YWG]-[QRFGS]-[SG]-[ESTAN]-G-x-
 [SAF].
- NAME: Glycosyl hydrolases family 1b active sites.
 25 CONSENSUS: E-[LIV]-D-[LIV]-x(0,1)-E-x(2)-[GQ]-[KRNf]-x-
 [PSTA].
- NAME: Glycosyl hydrolases family 17 signature.
 CONSENSUS: [LIVM]-x-[LIVMFYWA](3)-[STAG]-E-[STA]-G-W-P-[STN]-
 30 x-[SAGQ].
- NAME: Glycosyl hydrolases family 25 active sites signature.
 CONSENSUS: D-[LIVM]-x(3)-[NQ]-[PG]-x(9,10)-G-x(4)-
 [LIVMFY](2)-K-x-[ST]-E-[GS]-x(2)-
 35 CONSENSUS: Y-x-[DN].
- NAME: Glycosyl hydrolases family 31 active site.
 CONSENSUS: [GF]-[LIVMF]-W-x-D-M-[NSA]-E.
- 40 NAME: Glycosyl hydrolases family 31 signature 2.
 CONSENSUS: G-[AV]-D-[LIVMT]-C-G-[FY]-x(3)-[ST]-x(3)-L-C-x-R-
 W-x(2)-[LV]-[GS]-[SA]-
 CONSENSUS: F-x-P-F-x-R-[DN].
- 45 NAME: Glycosyl hydrolases family 32 active site.
 CONSENSUS: H-x(2)-P-x(4)-[LIVM]-N-D-P-N-G.
- NAME: Glycosyl hydrolases family 35 putative active site.
 50 CONSENSUS: G-G-P-[LIVM](2)-x(2)-Q-x-E-N-E-[FY].
- NAME: Glycosyl hydrolases family 39 active site.
 CONSENSUS: W-x-F-E-x-W-N-E-P-[DN].
- NAME: Glycosyl hydrolases family 45 active site.
 55 CONSENSUS: [STA]-T-R-Y-[FYW]-D-x(5)-[CA].
- NAME: Prokaryotic transglycosylases signature.

- CONSENSUS: [LIVM]-x(3)-E-S-x(3)-[AP]-x(3)-S-x(5)-G-[LIVM]-
 [LIVMFYW]-x-[LIVMFYW]-
 CONSENSUS: x(4)-[SAG].
- 5 NAME: Inosine-uridine preferring nucleoside hydrolase family
 signature.
 CONSENSUS: D-x-D-[PT]-[GA]-x-D-D-[TAV]-[VI]-A.
- 10 NAME: Alkylbase DNA glycosidases alkA family signature.
 CONSENSUS: G-I-G-x-W-[ST]-[AV]-x-[LIVMFY](2)-x-[LIVM]-x(8)-
 [MF]-x(2)-[ED]-D.
- 15 NAME: Formamidopyrimidine-DNA glycosylase signature.
 CONSENSUS: C-x(2,4)-C-x-[GTAQ]-x-[IV]-x(7)-R-[GSTAN]-[STA]-x-
 [FYI]-C-x(2)-C-Q.
- NAME: Uracil-DNA glycosylase signature.
 CONSENSUS: [KR]-[LIV]-[LIVC]-[LIVM]-x-G-[QI]-D-P-Y.
- 20 NAME: S-adenosyl-L-homocysteine hydrolase signature 1.
 CONSENSUS: [CS]-N-x-[FYL]-S-[ST]-[QA]-[DEN]-x-[AV](2)-A-A-
 [LIV]-[SAV].
- 25 NAME: S-adenosyl-L-homocysteine hydrolase signature 2.
 CONSENSUS: G-K-x(3)-[LIV]-x-G-Y-G-x-V-G-[KR]-G-x-A.
- NAME: Cytosol aminopeptidase signature.
 CONSENSUS: N-T-D-A-E-G-R-L.
- 30 NAME: Aminopeptidase P and proline dipeptidase signature.
 CONSENSUS: [HA]-[GSYR]-[LIVMT]-[SG]-H-x-[LIV]-G-[LIVM]-x-
 [IV]-H-[DE].
- 35 NAME: Methionine aminopeptidase subfamily 1 signature.
 CONSENSUS: [MFY]-x-G-H-G-[LIVMC]-[GSH]-x(3)-H-x(4)-[LIVM]-x-
 [HN]-[YWV].
- 40 NAME: Methionine aminopeptidase subfamily 2 signature.
 CONSENSUS: [DA]-[LIVMY]-x-K-[LIVM]-D-x-G-x-[HQ]-[LIVM]-[DNS]-
 G-x(3)-[DN].
- 45 NAME: Renal dipeptidase active site.
 CONSENSUS: [LIVM]-E-G-[GA]-x(2)-[LIVMF]-x(6)-L-x(3)-Y-x(2)-G-
 [LIVM]-R.
- NAME: Serine carboxypeptidases, serine active site.
 CONSENSUS: [LIVM]-x-[GTA]-E-S-Y-[AG]-[GS].
- 50 NAME: Serine carboxypeptidases, histidine active site.
 CONSENSUS: [LIVF]-x(2)-[LIVSTA]-x-[IVPST]-x-[GSDNQ]-[SAGV]-
 [SG]-H-x-[IVAQ]-P-x(3)-
 CONSENSUS: [PSA].
- 55 NAME: Zinc carboxypeptidases, zinc-binding region 1
 signature.
 CONSENSUS: [PK]-x-[LIVMFY]-x-[LIVMFY]-x(4)-H-[STAG]-x-E-x-
 [LIVM]-[STAG]-x(6)-
 CONSENSUS: [LIVMFYTA].

- NAME: Zinc carboxypeptidases, zinc-binding region 2
signature.
5 CONSENSUS: H-[ESTAG]-x(3)-[LIVME]-x(2)-[LIVMFYW]-P-[FYW].
- NAME: Serine proteases, trypsin family, histidine active site.
CONSENSUS: [LIVM]-[EST]-A-[ESTAG]-H-C.
- 10 NAME: Serine proteases, trypsin family, serine active site.
CONSENSUS: [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-
[SAPHV]-[LIVMFYWH]-
CONSENSUS: [LIVMFYSTANQH].
- 15 NAME: Serine proteases, subtilase family, aspartic acid active site.
CONSENSUS: [STAIV]-x-[LIVMF]-[LIVM]-D-[DSTA]-G-[LIVMFC]-
x(2,3)-[DNH].
- 20 NAME: Serine proteases, subtilase family, histidine active site.
CONSENSUS: H-G-[STM]-x-[VIC]-[ESTAGC]-[GS]-x-[LIVMA]-
[ESTAGCLV]-[SAGM].
- 25 NAME: Serine proteases, subtilase family, serine active site.
CONSENSUS: G-T-S-x-[SA]-x-P-x(2)-[STAVC]-[AG].
- NAME: Serine proteases, V8 family, histidine active site.
30 CONSENSUS: [ST]-G-[LIVMFYW](3)-[GN]-x(2)-T-[LIVM]-x-T-x(2)-H.
- NAME: Serine proteases, V8 family, serine active site.
CONSENSUS: T-x(2)-[GC]-[NQ]-S-G-S-x-[LIVM]-[FY].
- 35 NAME: Serine proteases, omptin family signature 1.
CONSENSUS: W-T-D-x-S-x-H-P-x-T.
- NAME: Serine proteases, omptin family signature 2.
40 CONSENSUS: A-G-Y-Q-E-[ST]-R-[FYW]-S-[FYW]-[TN]-A-x-G-G-[ST]-
Y.
- NAME: Prolyl endopeptidase family serine active site.
CONSENSUS: D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-
[LIVMFYW](2).
- 45 NAME: Endopeptidase Clp serine active site.
CONSENSUS: T-x(2)-[LIVMF]-G-x-A-[SAC]-S-[MSA]-[PAG]-[STA].
- NAME: Endopeptidase Clp histidine active site.
50 CONSENSUS: R-x(3)-[EAP]-x(3)-[LIVMFYT]-M-[LIVM]-H-Q-P.
- NAME: ATP-dependent serine proteases, lon family, serine active site.
CONSENSUS: D-G-[PD]-S-A-[GS]-[LIVMCA]-[TA]-[LIVM].
- 55 NAME: Eukaryotic thiol (cysteine) proteases cysteine active site.
CONSENSUS: Q-x(3)-[GE]-x-C-[YW]-x(2)-[ESTAGC]-[ESTAGCV].

NAME: Eukaryotic thiol (cysteine) proteases histidine active site.
 5 CONSENSUS: [LIVMGSTAN]-x-H-[GSACE]-[LIVM]-x-[LIVMAT](2)-G-x-[GSADNH].

NAME: Eukaryotic thiol (cysteine) proteases asparagine active site.
 10 CONSENSUS: [FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW]-
 CONSENSUS: [LIVMFYGG]-x-[LIVMF].

NAME: Ubiquitin carboxyl-terminal hydrolase family 1 cysteine active-site.
 15 CONSENSUS: Q-x(3)-N-[SA]-C-G-x(3)-[LIVM](2)-H-[SA]-[LIVM]-[SA].

NAME: Ubiquitin carboxyl-terminal hydrolases family 2 signature 1.
 20 CONSENSUS: G-[LIVMFY]-x(1,3)-[AGC]-[NASM]-x-C-[FYW]-[LIVMC]-[NST]-[SACV]-x-[LIVMS]-
 CONSENSUS: Q.

NAME: Ubiquitin carboxyl-terminal hydrolases family 2 signature 2.
 25 CONSENSUS: Y-x-L-x-[SAG]-[LIVMFT]-x(2)-H-x-G-x(4,5)-G-H-Y.

NAME: Caspase family histidine active site.
 30 CONSENSUS: H-x(2,4)-[SC]-x(4)-[LIVMF](2)-[ST]-H-G.

NAME: Caspase family cysteine active site.
 CONSENSUS: K-P-K-[LIVMF](4)-Q-A-C-[RQG]-G.

NAME: Eukaryotic and viral aspartyl proteases active site.
 35 CONSENSUS: [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-x-[LIVMFSTNC]-
 CONSENSUS: x-[LIVMFGTA].

NAME: Neutral zinc metallopeptidases, zinc-binding region signature.
 40 CONSENSUS: [GSTALIVN]-x(2)-H-E-[LIVMFYW]-[DEHRKP]-H-x-[LIVMFYWGSPQ].

NAME: Matrixins cysteine switch.
 45 CONSENSUS: P-R-C-[GN]-x-P-[DR]-[LIVSAPKQ].

NAME: Insulinase family, zinc-binding region signature.
 50 CONSENSUS: G-x(8,9)-G-x-[STA]-H-[LIVMFY]-[LIVMC]-[EDERN]-[HRKL]-[LMFAT]-x-[LFSTH]-x-
 CONSENSUS: [GSTAN]-[GST].

//

AC PS01016;
 55 DE Glycoprotease family signature.
 CONSENSUS: [KR]-[GSAT]-x(4)-[FYWHL]-[DQNGK]-x-P-x-[LIVMFY]-x(3)-H-x(2)-[AG]-H-
 CONSENSUS: [LIVM].

- NAME: Proteasome A-type subunits signature.
 5 CONSENSUS: [FY]-x(4)-[STNV]-x-[FYW]-S-P-x-G-[RKH]-x(2)-Q-[LIVM]-[DE]-Y-[SA]-x(2)-
 CONSENSUS: [SAG].
- NAME: Proteasome B-type subunits signature.
 10 CONSENSUS: [LIVMA]-[GSA]-[LIVMF]-x-[FYLVGAC]-x(2)-[GSACFY]-[LIVMSTAC](3)-[GAC]-
 CONSENSUS: [GSTACV]-[DES]-x(15)-[RK]-x(12,13)-G-x(2)-[GSTA]-D.
- NAME: Signal peptidases I serine active site.
 15 CONSENSUS: [GS]-x-S-M-x-[PS]-[AT]-[LF].
- NAME: Signal peptidases I lysine active site.
 CONSENSUS: K-R-[LIVMSTA](2)-G-x-[PG]-G-[DE]-x-[LIVM]-x-[LIVMFY].
- 20 NAME: Signal peptidases I signature 3.
 CONSENSUS: [LIVMFYW](2)-x(2)-G-D-[NH]-x(3)-[SND]-x(2)-[SG].
- NAME: Signal peptidases II signature.
 25 CONSENSUS: [GAF]-[GA]-[GAS]-[LIVM]-[GAS]-N-[LVMFG]-[LIVMFY]-D-R-[LIMFA].
- NAME: Peptidase family U32 signature.
 CONSENSUS: E-x-F-x(2)-G-[SA]-[LIVM]-C-x(4)-G-x-C-x-[LIVM]-S.
- 30 NAME: Amidases signature.
 CONSENSUS: G-[GA]-S-S-[GS]-G-x-[GSA]-[GSAVY]-x-[LIVM]-[GSA]-x(6)-[GSA]-x-[GA]-x-D-
 CONSENSUS: x-[GA]-x-S-[LIVM]-R-x-P-[GSAC].
- 35 NAME: Asparaginase / glutaminase active site signature 1.
 CONSENSUS: [LIVM]-x(2)-T-G-G-T-[IV]-[AGS].
- NAME: Asparaginase / glutaminase active site signature 2.
 40 CONSENSUS: G-x-[LIVM]-x(2)-H-G-T-D-T-[LIVM].
- NAME: Urease nickel ligands signature.
 CONSENSUS: T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P.
- NAME: Urease active site.
 45 CONSENSUS: [LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A.
- NAME: ArgE / dapE / ACY1 / CPG2 / yscS family signature 1.
 50 CONSENSUS: [LIV]-[GALMY]-[LIVMF]-x-[GSA]-H-x-D-[TV]-[STAV].
- NAME: ArgE / dapE / ACY1 / CPG2 / yscS family signature 2.
 CONSENSUS: [GSTAI]-[SANQ]-D-x-K-[GSACN]-x(2)-[LIVMA]-x(2)-[LIVMFY]-x(14,17)-[LIVM]-
 55 CONSENSUS: x-[LIVMF]-[LIVMSTAG]-[LIVMFA]-x(2)-[DNG]-E-E-x-[GSTN].
- NAME: Dihydroorotase signature 1.
 CONSENSUS: D-[LIVMFYWSAP]-H-[LIVA]-H-[LIVF]-[RN]-x-[PGN].

- NAME: Dihydroorotase signature 2.
 CONSENSUS: [GA]-[ST]-D-x-A-P-H-x(4)-K.
- 5 NAME: Beta-lactamase class-A active site.
 CONSENSUS: [FY]-x-[LIVMFY]-x-S-[TV]-x-K-x(4)-[AGLM]-x(2)-[LC].
- 10 NAME: Beta-lactamase class-C active site.
 CONSENSUS: F-E-[LIVM]-G-S-[LIVMG]-[SA]-K.
- NAME: Beta-lactamase class-D active site.
 CONSENSUS: [PA]-x-S-[ST]-F-K-[LIV]-[PAL]-x-[STA]-[LI].
- 15 NAME: Beta-lactamases class B signature 1.
 CONSENSUS: [LI]-x-[STN]-[HN]-x-H-[GSTA]-D-x(2)-G-[GP]-x(7,8)-[GS].
- 20 NAME: Beta-lactamases class B signature 2.
 CONSENSUS: P-x(3)-[LIVM](2)-x-G-x-C-[LIVMF](2)-K.
- NAME: Arginase family signature 1.
 CONSENSUS: [LIVMF]-G-G-x-H-x-[LIVMT]-[STAV]-x-[PAG]-x(3)-[GSTA].
- 25 NAME: Arginase family signature 2.
 CONSENSUS: [LIVM](2)-x-[LIVMFY]-D-[AS]-H-x-D.
- 30 NAME: Arginase family signature 3.
 CONSENSUS: [ST]-[LIVMFY]-D-[LIVM]-D-x(3)-[PAQ]-x(3)-P-[GSA]-x(7)-G.
- NAME: Adenosine and AMP deaminase signature.
 CONSENSUS: [SA]-[LIVM]-[NGS]-[STA]-D-D-P.
- 35 NAME: Cytidine and deoxycytidylate deaminases zinc-binding region signature.
 CONSENSUS: [CH]-[AGV]-E-x(2)-[LIVMFGAT]-[LIVM]-x(17,33)-P-C-x(2,8)-C-x(3)-[LIVM].
- 40 NAME: GTP cyclohydrolase I signature 1.
 CONSENSUS: [EN]-[LIVM](2)-x(2)-[KRQN]-[DN]-[LIVM]-x(3)-[ST]-x-C-E-H-H.
- 45 NAME: GTP cyclohydrolase I signature 2.
 CONSENSUS: [SA]-x-[RK]-x-Q-[LIVM]-Q-E-[RN]-[LI]-[TSN].
- NAME: Nitrilases / cyanide hydratase signature 1.
 CONSENSUS: G-x(2)-[LIVMFY](2)-x-[IF]-x-E-x(2)-[LIVM]-x-G-Y-P.
- 50 NAME: Nitrilases / cyanide hydratase active site signature.
 CONSENSUS: G-[GAQ]-x(2)-C-[WA]-E-[NH]-x(2)-[PST]-[LIVMFYS]-x-[KR].
- 55 NAME: Inorganic pyrophosphatase signature.
 CONSENSUS: D-[SGDN]-D-[PE]-[LIVMF]-D-[LIVMGAC].
- NAME: Acylphosphatase signature 1.

- CONSENSUS: [LIV]-x-G-x-V-Q-G-V-x-[FM]-R.
- NAME: Acylphosphatase signature 2.
 5 CONSENSUS: G-[FYW]-[AVC]-[KRQAM]-N-x(3)-G-x-V-x(5)-G.
- NAME: ATP synthase alpha and beta subunits signature.
 CONSENSUS: P-[SAP]-[LIV]-[DNH]-x(3)-S-x-S.
- NAME: ATP synthase gamma subunit signature.
 10 CONSENSUS: [IV]-T-x-E-x(2)-[DE]-x(3)-G-A-x-[SAKR].
- NAME: ATP synthase delta (OSCP) subunit signature.
 CONSENSUS: [LIVM]-x-[LIVMFYT]-x(3)-[LIVMT]-[DENQK]-x(2)-
 [LIVM]-x-[GSA]-G-[LIVMFYGA]-
 15 CONSENSUS: x-[LIVM]-[KRHENQ]-x-[GSEN].
- NAME: ATP synthase a subunit signature.
 CONSENSUS: [STAGN]-x-[STAG]-[LIVMF]-R-L-x-[SAGV]-N-[LIVMT].
- NAME: ATP synthase c subunit signature.
 20 CONSENSUS: [GSTA]-R-[NQ]-P-x(10)-[LIVMFYW](2)-x(3)-[LIVMFYW]-
 x-[DE].
- NAME: E1-E2 ATPases phosphorylation site.
 25 CONSENSUS: D-K-T-G-T-[LI]-[TI].
- NAME: Sodium and potassium ATPases beta subunits signature
 1.
 CONSENSUS: [FYW]-x(2)-[FYW]-x-[FYW]-[DN]-x(6)-[LIVM]-G-R-T-
 30 x(3)-W.
- NAME: Sodium and potassium ATPases beta subunits signature
 2.
 CONSENSUS: [RK]-x(2)-C-[RKQWI]-x(5)-L-x(2)-C-[SA]-G.
 35
- NAME: GDA1/CD39 family of nucleoside phosphatases signature.
 CONSENSUS: [LIVM]-x-G-x(2)-E-G-x-[FY]-x-[FW]-[LIVA]-[TAG]-x-
 N-[HY].
- NAME: Iodothyronine deiodinases active site.
 40 CONSENSUS: R-P-L-V-x-N-F-G-S-[CA]-T-C-P-x-F.
- NAME: Cutinase, serine active site.
 CONSENSUS: P-x-[STA]-x-[LIV]-[IVT]-x-[GS]-G-Y-S-[QL]-G.
 45
- NAME: Cutinase, aspartate and histidine active sites.
 CONSENSUS: C-x(3)-D-x-[IV]-C-x-G-[GST]-x(2)-[LIVM]-x(2,3)-H.
- NAME: DDC / GAD / HDC / TyrDC pyridoxal-phosphate attachment
 50 site.
 CONSENSUS: S-[LIVMFYW]-x(5)-K-[LIVMFYWG](2)-x(3)-[LIVMFYW]-x-
 [CA]-x(2)-[LIVMFYWQ]-
 CONSENSUS: x(2)-[RK].
- NAME: Orn/Lys/Arg decarboxylases family 1 pyridoxal-P
 55 attachment site.
 CONSENSUS: [STAV]-x-S-x-H-K-x(2)-[GSTAN](2)-x-[STA]-Q-
 [STA](2).

- NAME: Orn/DAP/Arg decarboxylases family 2 pyridoxal-P attachment site.
 5 CONSENSUS: [FY]-[PA]-x-K-[SACV]-[NHCLFW]-x(4)-[LIVMF]-[LIVMTA]-x(2)-[LIVMA]-x(3)-
 CONSENSUS: [GTE].
- NAME: Orn/DAP/Arg decarboxylases family 2 signature 2.
 10 CONSENSUS: [GS]-x(2,6)-[LIVMSCP]-x(2)-[LIVMF]-[DNS]-[LIVMCA]-G-G-G-[LIVMFY]-
 CONSENSUS: [GSTPCEQ].
- NAME: Orotidine 5'-phosphate decarboxylase active site.
 15 CONSENSUS: [LIVMFTA]-[LIVMF]-x-D-x-K-x(2)-D-I-[GP]-x-T-[LIVMTA].
- NAME: Phosphoenolpyruvate carboxylase active site 1.
 CONSENSUS: [VT]-x-T-A-H-P-T-[EQ]-x(2)-R-[KRH].
- 20 NAME: Phosphoenolpyruvate carboxylase active site 2.
 CONSENSUS: [IV]-M-[LIVM]-G-Y-S-D-S-x-K-D-[STAG]-G.
- NAME: Phosphoenolpyruvate carboxykinase (GTP) signature.
 25 CONSENSUS: F-P-S-A-C-G-K-T-N.
- NAME: Phosphoenolpyruvate carboxykinase (ATP) signature.
 CONSENSUS: L-I-G-D-D-E-H-x-W-x-[DE]-x-G-[IV]-x-N.
- 30 NAME: Uroporphyrinogen decarboxylase signature 1.
 CONSENSUS: P-x-W-x-M-R-Q-A-G-R.
- NAME: Uroporphyrinogen decarboxylase signature 2.
 35 CONSENSUS: G-F-[STAGCV]-[STAGC]-x-P-[FYW]-T-[LV]-x(2)-Y-x(2)-[AE]-[GK].
- NAME: Indole-3-glycerol phosphate synthase signature.
 CONSENSUS: [LIVMFY]-[LIVMC]-x-E-[LIVMFYC]-K-[KRSP]-[STAK]-S-P-[ST]-x(3)-[LIVMFYST].
- 40 NAME: Ribulose biphosphate carboxylase large chain active site.
 CONSENSUS: G-x-[DN]-F-x-K-x-D-E.
- NAME: Fructose-bisphosphate aldolase class-I active site.
 45 CONSENSUS: [LIVM]-x-[LIVMFYW]-E-G-x-[LS]-L-K-P-[SN].
- NAME: Fructose-bisphosphate aldolase class-II signature 1.
 CONSENSUS: [FYVM]-x(1,3)-[LIVMH]-[APN]-[LIVM]-x(1,2)-[LIVM]-H-x-D-H-[GACH].
- 50 NAME: Fructose-bisphosphate aldolase class-II signature 2.
 CONSENSUS: [LIVM]-E-x-E-[LIVM]-G-x(2)-[GM]-[GSTA]-x-E.
- NAME: Malate synthase signature.
 55 CONSENSUS: [KR]-[DENQ]-H-x(2)-G-L-N-x-G-x-W-D-Y-[LIVM]-F.
- NAME: Hydroxymethylglutaryl-coenzyme A lyase active site.
 CONSENSUS: S-V-A-G-L-G-G-C-P-Y.

- NAME: Hydroxymethylglutaryl-coenzyme A synthase active site.
 CONSENSUS: N-x-[DN]-[IV]-E-G-[IV]-D-x(2)-N-A-C-[FY]-x-G.
- 5 NAME: Citrate synthase signature.
 CONSENSUS: G-[FYA]-[GA]-H-x-[IV]-x(1,2)-[RKT]-x(2)-D-[PS]-R.
- NAME: Alpha-isopropylmalate and homocitrate synthases signature 1.
 10 CONSENSUS: L-R-[DE]-G-x-Q-x(10)-K.
- NAME: Alpha-isopropylmalate and homocitrate synthases signature 2.
 CONSENSUS: [LIVMF]-x(2)-H-x-H-[DN]-D-x-G-x-[GAS]-x-[GASLI].
- 15 NAME: KDPG and KHG aldolases active site.
 CONSENSUS: G-[LIVM]-x(3)-E-[LIV]-T-[LF]-R.
- NAME: KDPG and KHG aldolases Schiff-base forming residue.
 20 CONSENSUS: G-x(3)-[LIVMF]-K-[LF]-F-P-[SA]-x(3)-G.
- NAME: Isocitrate lyase signature.
 CONSENSUS: K-[KR]-C-G-H-[LMQ].
- 25 NAME: Beta-eliminating lyases pyridoxal-phosphate attachment site.
 CONSENSUS: Y-x-D-x(3)-M-S-[GA]-K-K-D-x-[LIVM](2)-x-[LIVM]-G-G.
- 30 NAME: DNA photolyases class 1 signature 1.
 CONSENSUS: T-G-x-P-[LIVM](2)-D-A-x-M-[RA]-x-[LIVM].
- NAME: DNA photolyases class 1 signature 2.
 CONSENSUS: [DN]-R-x-R-[LIVM](2)-x-[STA](2)-F-[LIVMFA]-x-K-x-L-x(2,3)-W-[KRQ].
- 35 NAME: DNA photolyases class 2 signature 1.
 CONSENSUS: F-x-E-E-x-[LIVM](2)-R-R-E-L-x(2)-N-F.
- 40 NAME: DNA photolyases class 2 signature 2.
 CONSENSUS: G-x-H-D-x(2)-W-x-E-R-x-[LIVM]-F-G-K-[LIVM]-R-[FY]-M-N.
- NAME: Eukaryotic-type carbonic anhydrases signature.
 45 CONSENSUS: S-E-H-x-[LIVM]-x(4)-[FYH]-x(2)-E-[LIVM]-H-[LIVMFA](2).
- NAME: Prokaryotic-type carbonic anhydrases signature 1.
 CONSENSUS: C-[SA]-D-S-R-[LIVM]-x-[AP].
- 50 NAME: Prokaryotic-type carbonic anhydrases signature 2.
 CONSENSUS: [EQ]-Y-A-[LIVM]-x(2)-[LIVM]-x(4)-[LIVMF](3)-x-G-H-x(2)-C-G.
- 55 NAME: Fumarate lyases signature.
 CONSENSUS: G-S-x(2)-M-x(2)-K-x-N.
- NAME: Aconitase family signature 1.

CONSENSUS: [LIVM]-x(2)-[GSACIVM]-x-[LIV]-[GTIV]-[STP]-C-
x(D,L)-T-N-[GSTANI]-x(4)-
CONSENSUS: [LIVMA].

5 NAME: Aconitase family signature 2.
CONSENSUS: G-x(2)-[LIVWPQ]-x(3)-[GAC]-C-[GSTAM]-[LIMPTA]-C-
[LIMV]-[GA].

10 NAME: Dihydroxy-acid and b-phosphogluconate dehydratases
signature 1.
CONSENSUS: C-D-K-x(2)-P-[GA]-x(3)-[GA].

NAME: Dihydroxy-acid and b-phosphogluconate dehydratases
signature 2.
15 CONSENSUS: [SA]-L-[LIVM]-T-D-[GA]-R-[LIVMF]-S-[GA]-[GAV]-
[ST].

NAME: Dehydroquinase class I active site.
20 CONSENSUS: D-[LIVM]-[DE]-[LIVN]-x(18,20)-[LIVM](2)-x-[SC]-
[NHY]-H-[DN].

NAME: Dehydroquinase class II signature.
25 CONSENSUS: [LIVM]-[NQ]-G-P-N-[LV]-x(2)-L-G-x-R-[QED]-P-x(2)-
[FY]-G.

NAME: Enolase signature.
CONSENSUS: [LIV](3)-K-x-N-Q-I-G-[ST]-[LIV]-[ST]-[DE]-[STA].

30 NAME: Serine/threonine dehydratases pyridoxal-phosphate
attachment site.
CONSENSUS: [DESH]-x(4,5)-[STVG]-x-[AS]-[FYI]-K-[DLIFSA]-
[RVMF]-[GA]-[LIVMGA].

NAME: Enoyl-CoA hydratase/isomerase signature.
35 CONSENSUS: [LIVM]-[STA]-x-[LIVM]-[DENQRHSTA]-G-x(3)-[AG](3)-
x(4)-[LIVMST]-x-[CSTA]-
CONSENSUS: [DQHP]-[LIVMFY].

NAME: Imidazoleglycerol-phosphate dehydratase signature 1.
40 CONSENSUS: [LIVMY]-[DE]-x-H-H-x(2)-E-x(2)-[GCA]-[LIVM]-
[STAC]-[LIVM].

NAME: Imidazoleglycerol-phosphate dehydratase signature 2.
45 CONSENSUS: G-x-[DN]-x-H-H-x(2)-E-[STAGC]-x-[FY]-K.

NAME: Tryptophan synthase alpha chain signature.
CONSENSUS: [LIVM]-E-[LIVM]-G-x(2)-[FYC]-[ST]-[DE]-[PA]-
[LIVMY]-[AGLI]-[DE]-G.

50 NAME: Tryptophan synthase beta chain pyridoxal-phosphate
attachment site.
CONSENSUS: [LIVM]-x-H-x-G-[STA]-H-K-x-N.

NAME: Delta-aminolevulinic acid dehydratase active site.
55 CONSENSUS: G-x-D-x-[LIVM](2)-[IV]-K-P-[GSA]-x(2)-Y.

NAME: Urocanase active site.
CONSENSUS: F-Q-G-L-P-x-R-I-C-W.

- NAME: Prephenate dehydratase signature 1.
 5 CONSENSUS: [FY]-x-[LIVM]-x(2)-[LIVM]-x(5)-[DN]-x(5)-T-R-F-[LIVM]-x-[LIVM].
- NAME: Prephenate dehydratase signature 2.
 CONSENSUS: [LIVM]-[ST]-[KR]-[LIVM]-E-[ST]-R-P.
- NAME: Dihydrodipicolinate synthetase signature 1.
 10 CONSENSUS: [GSA]-[LIVM]-[LIVMFY]-x(2)-G-[ST]-[TG]-G-E-[GASNF]-x(6)-[EQ].
- NAME: Dihydrodipicolinate synthetase signature 2.
 15 CONSENSUS: Y-[DNS]-[LIVMF]-P-x(2)-[ST]-x(3)-[LIVM]-x(13,14)-[LIVM]-x-[SGA]-[LIVMF]-
 CONSENSUS: K-[DEQAF]-[STAC].
- NAME: RsaA family of pseudouridine synthase signature.
 20 CONSENSUS: G-R-L-D-x(2)-[ST]-x-G-[LIVMF](4)-[ST]-[DNT].
- NAME: Cysteine synthase/cystathionine beta-synthase P-phosphate attachment site.
 25 CONSENSUS: K-x-E-x(3)-[PA]-[STAGC]-x-S-[IVAP]-K-x-R-x-[STAG]-x(2)-[LIVM].
- NAME: Phenylalanine and histidine ammonia-lyases signature.
 CONSENSUS: G-[STG]-[LIVM]-[STG]-[AC]-S-G-[DH]-L-x-P-L-[SA]-x(2)-[SA].
- NAME: Porphobilinogen deaminase cofactor-binding site.
 30 CONSENSUS: E-R-x-[LIVMFA]-x(3)-[LIVMF]-x-G-[GSA]-C-x-[IVT]-P-[LIVMF]-[GSA].
- NAME: Cys/Met metabolism enzymes pyridoxal-phosphate attachment site.
 35 CONSENSUS: [DQ]-[LIVMF]-x(3)-[STAGC]-[STAGCI]-T-K-[FYWQ]-[LIVMF]-x-G-[HQ]-[SGNH].
- NAME: Glyoxalase I signature 1.
 40 CONSENSUS: [HQ]-[IVT]-x-[LIVFY]-x-[IV]-x(5)-[STA]-x(2)-F-[YM]-x(2,3)-[LMF]-G-[LMF].
- NAME: Glyoxalase I signature 2.
 45 CONSENSUS: G-[ENTKQ]-x(0,5)-[GA]-[LVFY]-[GH]-H-[IVF]-[CGA]-x-[STAGL]-x(2)-[DNC].
- NAME: Cytochrome c and c1 heme lyases signature 1.
 CONSENSUS: H-N-x(2)-N-E-x(2)-W-[NQKR]-x(4)-W-E.
- NAME: Cytochrome c and c1 heme lyases signature 2.
 50 CONSENSUS: P-F-D-R-H-D-W.
- NAME: Adenylate cyclases class-I signature 1.
 CONSENSUS: E-Y-F-G-[SA](2)-L-W-x-L-Y-K.
- NAME: Adenylate cyclases class-I signature 2.
 55 CONSENSUS: Y-R-N-x-W-[NS]-E-[LIVM]-R-T-L-H-F-x-G.

- NAME: Guanylate cyclases signature.
 CONSENSUS: G-V-[LIVM]-x(0,1)-G-x(5)-[FY]-x-[LIVM]-[FYW]-[GS]-
 [DNTHKW]-[DNT]-[IV]-
 CONSENSUS: [DNTA]-x(5)-[DE].
- 5 NAME: Chorismate synthase signature 1.
 CONSENSUS: G-E-S-H-[GC]-x(2)-[LIVM]-[GTV]-x-[LIVM](2)-[DE]-G-
 x-[PV].
- 10 NAME: Chorismate synthase signature 2.
 CONSENSUS: [GE]-R-[SA](2)-[SAG]-R-[EV]-[ST]-x(2)-[RH]-V-x(2)-
 G.
- 15 NAME: Chorismate synthase signature 3.
 CONSENSUS: R-[SH]-D-[PSV]-[CSAV]-x(4)-[GAI]-x-[IVGSP]-[LIVM]-
 x-E-[STAH]-[LIVM].
- NAME: L-pyruvoyl tetrahydropterin synthase signature 1.
 CONSENSUS: C-N-N-x(2)-G-H-G-H-N-Y.
- 20 NAME: L-pyruvoyl tetrahydropterin synthase signature 2.
 CONSENSUS: D-H-K-N-L-D-x-D.
- NAME: Ferrochelatase signature.
 CONSENSUS: [LIVMF](2)-x-S-x-H-[GS]-[LIVM]-P-x(4,5)-[DENQKR]-
 x-G-D-x-Y.
- 25 NAME: Alanine racemase pyridoxal-phosphate attachment site.
 CONSENSUS: V-x-K-A-[DN]-[GA]-Y-G-H-G.
- 30 NAME: Aspartate and glutamate racemases signature 1.
 CONSENSUS: [IVA]-[LIVM]-x-C-x(0,1)-N-[ST]-[MSA]-[STH]-
 [LIVFYSTANK].
- 35 NAME: Aspartate and glutamate racemases signature 2.
 CONSENSUS: [LIVM](2)-x-[AG]-C-T-[DEH]-[LIVMFY]-[PNGRS]-x-
 [LIVM].
- NAME: Mandelate racemase / muconate lactonizing enzyme
 family signature 1.
 CONSENSUS: A-x-[SAG](2)-[LIVM]-[DE]-x-A-x(2)-D-x(2)-[GA]-
 [KR].
- 40 NAME: Mandelate racemase / muconate lactonizing enzyme
 family signature 2.
 CONSENSUS: G-x(7)-D-x(9)-A-x(14)-[LIVM]-E-[DENQ]-P-x(4)-
 [DENQ].
- 45 NAME: Ribulose-phosphate 3-epimerase family signature 1.
 CONSENSUS: [LIVMF]-H-[LIVMFY]-D-[LIVM]-x-D-x(1,2)-[FY]-
 [LIVM]-x-N-x-[STAV].
- 50 NAME: Ribulose-phosphate 3-epimerase family signature 2.
 CONSENSUS: [LIVMA]-x-[LIVM]-M-[ST]-[VS]-x-P-x(3)-G-Q-x-F-
 x(6)-[NK]-[LIVMC].
- 55 NAME: Aldose 1-epimerase putative active site.
 CONSENSUS: [NS]-x-T-N-H-x-Y-[FW]-N-[LI].

- NAME: Cyclophilin-type peptidyl-prolyl cis-trans isomerase signature.
 5 CONSENSUS: [FY]-x(2)-[STCNLV]-x-F-H-[RH]-[LIVMN]-[LIVM]-x(2)-F-[LIVM]-x-Q-[AG]-G.
- NAME: Cyclophilin-type peptidyl-prolyl cis-trans isomerase profile.
- 10 NAME: FKBP-type peptidyl-prolyl cis-trans isomerase signature 1.
 CONSENSUS: [LIVMC]-x-[YF]-x-[GVL]-x(1,2)-[LFT]-x(2)-G-x(3)-[DE]-[STAEQK]-[STAN].
- 15 NAME: FKBP-type peptidyl-prolyl cis-trans isomerase signature 2.
 CONSENSUS: [LIVMFY]-x(2)-[GA]-x(3,4)-[LIVMF]-x(2)-[LIVMFHK]-x(2)-G-x(4)-[LIVMF]-
 CONSENSUS: x(3)-[PSGAQ]-x(2)-[AG]-[FY]-G.
- 20 NAME: FKBP-type peptidyl-prolyl cis-trans isomerase domain profile.
- 25 NAME: PpiC-type peptidyl-prolyl cis-trans isomerase signature.
 CONSENSUS: F-[GSADEI]-x-[LVAQ]-A-x(3)-[ST]-x(3,4)-[STQ]-x(3,5)-[GER]-G-x-[LIVM]-
 CONSENSUS: [GS].
- 30 NAME: Triosephosphate isomerase active site.
 CONSENSUS: [AV]-Y-E-P-[LIVM]-W-[SA]-I-G-T-[GK].
- NAME: Xylose isomerase signature 1.
 35 CONSENSUS: [LI]-E-P-K-P-x(2)-P.
- NAME: Xylose isomerase signature 2.
 CONSENSUS: [FL]-H-D-x-D-[LIV]-x-[PD]-x-[GDE].
- NAME: Phosphomannose isomerase type I signature 1.
 40 CONSENSUS: Y-x-D-x-N-H-K-P-E.
- NAME: Phosphomannose isomerase type I signature 2.
 CONSENSUS: H-A-Y-[LIVM]-x-G-x(2)-[LIVM]-E-x-M-A-x-S-D-N-x-[LIVM]-R-A-G-x-T-P-K.
- 45 NAME: Phosphoglucose isomerase signature 1.
 CONSENSUS: [DENS]-x-[LIVM]-G-G-R-[FY]-S-[LIVMT]-x-[STA]-[PSAC]-[LIVMA]-G.
- NAME: Phosphoglucose isomerase signature 2.
 50 CONSENSUS: [GS]-x-[LIVM]-[LIVMFYW]-x(4)-[FY]-[DN]-Q-x-G-V-E-x(2)-K.
- NAME: Glucosamine/galactosamine-6-phosphate isomerases signature.
 55 CONSENSUS: [LIVM]-x(3)-G-x-[LIT]-x-[LIV]-x-[LIVM]-x-G-[LIVM]-G-x-[DEN]-G-H.

- NAME: Phosphoglycerate mutase family phosphohistidine
signature.
CONSENSUS: [LIVM]-x-R-H-G-[EQ]-x(3)-N.
- 5 NAME: Phosphoglucumutase and phosphomannomutase
phosphoserine signature.
CONSENSUS: [GSA]-[LIVM]-x-[LIVM]-[ST]-[PGA]-S-H-x-P-x(4)-
[GNHE].
- 10 NAME: Methylmalonyl-CoA mutase signature.
CONSENSUS: R-I-A-R-N-[TQ]-x(2)-[LIVMFY](2)-x-[EQ]-E-x(4)-
[KRN]-x(2)-D-P-x-[GSA]-
CONSENSUS: G-S.
- 15 NAME: Terpene synthases signature.
CONSENSUS: [DE]-G-S-W-x-G-x-W-[GA]-[LIVM]-x-[FY]-x-Y-[GA].
- NAME: Eukaryotic DNA topoisomerase I active site.
CONSENSUS: [DEN]-x(6)-[GS]-[IT]-S-K-x(2)-Y-[LIVM]-x(3)-
20 [LIVM].
- NAME: Prokaryotic DNA topoisomerase I active site.
CONSENSUS: [EQ]-x-L-Y-[DEQT]-x(3,12)-[LI]-[ST]-Y-x-R-[ST]-
[DEQS].
- 25 NAME: DNA topoisomerase II signature.
CONSENSUS: [LIVMA]-x-E-G-[DN]-S-A-x-[STAG].
- NAME: Aminoacyl-transfer RNA synthetases class-I signature.
30 CONSENSUS: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-
[LIVMYAC]-G-[HNTG]-
CONSENSUS: [LIVMFYSTAGPC].
- NAME: Aminoacyl-transfer RNA synthetases class-II signature
1.
35 CONSENSUS: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE].
- NAME: Aminoacyl-transfer RNA synthetases class-II signature
2.
40 CONSENSUS: [GSTALVF]-[DENQHRKP]-[GSTA]-[LIVMF]-[DE]-R-
[LIVMF]-x-[LIVMSTAG]-[LIVMFY].
- NAME: WHEP-TRS domain signature.
CONSENSUS: [QY]-G-[DNEA]-x-[LIV]-[KR]-x(2)-K-x(2)-[KRNG]-
45 [AS]-x(4)-[LIV]-[DENK]-
CONSENSUS: x(2)-[IV]-x(2)-L-x(3)-K.
- NAME: ATP-citrate lyase / succinyl-CoA ligases family
signature 1.
50 CONSENSUS: S-[KR]-S-G-[GT]-[LIVM]-[GST]-x-[EQ]-x(8,10)-G-
x(4)-[LIVM]-[GA]-[LIVM]-G-
CONSENSUS: G-D.
- NAME: ATP-citrate lyase / succinyl-CoA ligases family active
55 site.
CONSENSUS: G-x(2)-A-x(4,7)-[RQT]-[LIVMF]-G-H-[AS]-[GH].

NAME: ATP-citrate lyase / succinyl-CoA ligases family
signature 3.
CONSENSUS: G-x-[IV]-x(2)-[LIVMF]-x-[NA]-G-[GA]-G-[LA]-[STAV]-
x(4)-D-x-[LIVM]-x(3)-
5 CONSENSUS: G-[GRE].

NAME: Glutamine synthetase signature 1.
CONSENSUS: [FYWL]-D-G-S-S-x(6,8)-[DENQSTAK]-[SA]-[DE]-x(2)-
[LIVMFY].

10 NAME: Glutamine synthetase putative ATP-binding region
signature.
CONSENSUS: K-P-[LIVMFYA]-x(3,5)-[NPAT]-G-[GSTAN]-G-x-H-x(3)-
S.

15 NAME: Glutamine synthetase class-I adenylation site.
CONSENSUS: K-[LIVM]-x(5)-[LIVMA]-D-[RK]-[DN]-[LI]-Y.

20 NAME: D-alanine--D-alanine ligase signature 1.
CONSENSUS: H-G-x(2)-G-E-D-G-x-[LIVMA]-[QSA]-[GSA].

NAME: D-alanine--D-alanine ligase signature 2.
CONSENSUS: [LIV]-x(3)-[GA]-x-[GSAIV]-R-[LIVCA]-D-[LIVMF](2)-
x(7,9)-[LI]-x-E-
25 CONSENSUS: [LIVA]-N-[STP]-x-P-[GA].

NAME: SAICAR synthetase signature 1.
CONSENSUS: [LIVMF](2)-P-[LIVM]-E-x-[LIVM]-[LIVMCA]-R-x(3)-
[TA]-G-S.

30 NAME: SAICAR synthetase signature 2.
CONSENSUS: [LIVM]-[LIVMA]-D-x-K-[LIVMFY]-E-F-G.

35 NAME: Folylpolyglutamate synthase signature 1.
CONSENSUS: [LIVMFY]-x-[LIVM]-[STAG]-G-T-[NK]-G-K-x-[ST]-x(7)-
[LIVM](2)-x(3)-[GSK].

NAME: Folylpolyglutamate synthase signature 2.
CONSENSUS: [LIVMFY](2)-E-x-G-[LIVM]-[GA]-G-x(2)-D-x-[GST]-x-
40 [LIVM](2).

NAME: Ubiquitin-activating enzyme signature 1.
CONSENSUS: K-A-C-S-G-K-F-x-P.

45 NAME: Ubiquitin-activating enzyme active site.
CONSENSUS: P-[LIVM]-C-T-[LIVM]-[KRH]-x-[FT]-P.

NAME: Ubiquitin-conjugating enzymes active site.
CONSENSUS: [FYWLSP]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-
50 [LIV]-x-[LIV].

NAME: Formate--tetrahydrofolate ligase signature 1.
CONSENSUS: G-[LIVM]-K-G-G-A-A-G-G-G-Y.

55 NAME: Formate--tetrahydrofolate ligase signature 2.
CONSENSUS: V-A-T-[IV]-R-A-L-K-x-[HN]-G-G.

NAME: Adenylosuccinate synthetase GTP-binding site.

CONSENSUS: Q-W-G-D-E-G-K-G.

NAME: Adenylosuccinate synthetase active site.

CONSENSUS: G-I-[GR]-P-x-Y-x(2)-K-x(2)-R.

5

NAME: Argininosuccinate synthase signature 1.

CONSENSUS: A-[FY]-S-G-G-L-D-T-S.

NAME: Argininosuccinate synthase signature 2.

CONSENSUS: G-x-T-x-K-G-N-D-x(2)-R-F.

10

NAME: Phosphoribosylglycinamide synthetase signature.

CONSENSUS: R-F-G-D-P-E-x-[QM].

NAME: Carbamoyl-phosphate synthase subdomain signature 1.

CONSENSUS: [FYV]-[PS]-[LIVMC]-[LIVMA]-[LIVM]-[KR]-[PSA]-[STA]-x(3)-[SG]-G-x-[AG].

15

NAME: Carbamoyl-phosphate synthase subdomain signature 2.

CONSENSUS: [LIVMF]-[LIMN]-E-[LIVMA]-N-[PATLIVM]-[KR]-[LIVMSTAC].

20

NAME: ATP-dependent DNA ligase AMP-binding site.

CONSENSUS: [EDQH]-x-K-x-[DN]-G-x-R-[GACIVM].

25

NAME: ATP-dependent DNA ligase signature 2.

CONSENSUS: E-G-[LIVMA]-[LIVM](2)-[KR]-x(5,8)-[YW]-[QNEK]-x(2,6)-[KRH]-x(3,5)-K-

CONSENSUS: [LIVMFY]-K.

30

NAME: NAD-dependent DNA ligase signature 1.

CONSENSUS: K-[LIVM]-D-G-[LIVM]-[SA]-x(4)-Y-x(2)-G-x-L-x(4)-[ST]-R-G-[DN]-G-x(2)-G-

CONSENSUS: [DE]-[DENL].

35

NAME: NAD-dependent DNA ligase signature 2.

CONSENSUS: [IV]-G-[KR]-[ST]-G-x-[LIVM]-[STNK]-x-[VT]-x(2)-L-x-[PS]-V.

NAME: RNA 3'-terminal phosphate cyclase signature.

CONSENSUS: [RH]-G-x(2)-P-x-G(3)-x-[LIV].

40

NAME: Lipoate-protein ligase B signature.

CONSENSUS: R-G-G-x(2)-T-[FYW]-H-x(2)-[GH]-Q-x-[LIV]-x-Y.

45

NAME: Isopenicillin N synthetase signature 1.

CONSENSUS: [RK]-x-[STA]-x(2)-S-x-C-Y-[SL].

NAME: Isopenicillin N synthetase signature 2.

CONSENSUS: [LIVM](2)-x-C-G-[STA]-x(2)-[STAG]-x(2)-T-x-[DNG].

50

NAME: Site-specific recombinases active site.

CONSENSUS: Y-[LIVAC]-R-[VA]-S-[ST]-x(2)-Q.

NAME: Site-specific recombinases signature 2.

CONSENSUS: G-[DE]-x(2)-[LIVM]-x(3)-[LIVM]-[DT]-R-[LIVM]-[GSA].

55

- NAME: Transposases, Mutator family, signature.
 CONSENSUS: D-x(3)-G-[LIVMF]-x(6)-[STAV]-[LIVMFYW]-[PT]-x-
 [STAV]-x(2)-[QR]-x-C-x(2)-
 CONSENSUS: H.
- 5 NAME: Transposases, IS30 family, signature.
 CONSENSUS: R-G-x(2)-E-N-x-N-G-[LIVM](2)-R-[QE]-[LIVMFY](2)-P-
 K.
- 10 NAME: Autoinducers synthetases family signature.
 CONSENSUS: [LMFY]-R-x(3)-F-x(2)-[KR]-x(2)-W-x-[LIVM]-x(6,9)-
 E-x-D-x-[FY]-D.
- 15 NAME: Thiamine pyrophosphate enzymes signature.
 CONSENSUS: [LIVMF]-[GSA]-x(5)-P-x(4)-[LIVMFYW]-x-[LIVMF]-x-G-
 D-[GSA]-[GSAC].
- NAME: Biotin-requiring enzymes attachment site.
 CONSENSUS: [GN]-[DEQTR]-x-[LIVMFY]-x(2)-[LIVM]-x-[AIV]-M-K-
 20 [LMAT]-x(3)-[LIVM]-x-
 CONSENSUS: [SAV].
- NAME: 2-oxo acid dehydrogenases acyltransferase component
 lipoyl binding site.
 25 CONSENSUS: [GN]-x(2)-[LIVF]-x(5)-[LIVFC]-x(2)-[LIVFA]-x(3)-K-
 [STAIV]-[STAVQDN]-
 CONSENSUS: x(2)-[LIVMFS]-x(5)-[GCN]-x-[LIVMFY].
- NAME: Putative AMP-binding domain signature.
 30 CONSENSUS: [LIVMFY]-x(2)-[STG]-[STAG]-G-[ST]-[STEI]-[SG]-x-
 [PASLIVM]-[KR].
- NAME: Molybdenum cofactor biosynthesis proteins signature 1.
 CONSENSUS: [LIVM](3)-[LIT](2)-G-G-T-G-x(4)-D.
- 35 NAME: Molybdenum cofactor biosynthesis proteins signature 2.
 CONSENSUS: S-x-[GS]-x(2)-D-x(5)-[LIVW]-x(10,12)-[LIV]-x(2)-
 [KR]-P-G-[KRL]-P-x(2)-
 CONSENSUS: [LIVMF]-[GA].
- 40 NAME: moaA / nifB / pqqE family signature.
 CONSENSUS: [LIV]-x(3)-C-[NP]-[LIVMF]-[QRS]-C-x-[FYM]-C.
- NAME: Radical activating enzymes signature.
 45 CONSENSUS: [GV]-x-G-x-[KR]-x(3)-F-x(2)-G-x(0,1)-C-x(3)-C-
 x(2)-C-x-[NL].
- NAME: Tpx family signature.
 CONSENSUS: S-x-D-L-P-F-A-x(2)-[KR]-[FW]-C.
- 50 NAME: Cytochrome c family heme-binding site signature.
 CONSENSUS: C-{CPWHF}-{CPWR}-C-H-{CFYW}.
- NAME: Cytochrome b5 family, heme-binding domain signature.
 55 CONSENSUS: [FY]-[LIVMK]-x(2)-H-P-[GA]-G.
- NAME: Cytochrome b/b6 heme-ligand signature.
 CONSENSUS: [DENQ]-x(3)-G-[FYWMQ]-x-[LIVMF]-R-x(2)-H.

- NAME: Cytochrome b/bb 20 site signature.
 CONSENSUS: P-[DE]-W-[FY]-[LFY](2).
- 5 NAME: Cytochrome b559 subunits heme-binding site signature.
 CONSENSUS: [LIV]-x-[ST]-[LIVF]-R-[FYW]-x(2)-[IV]-H-[STGA]-[LIV]-[STGA]-[IV]-P.
- 10 NAME: Nickel-dependent hydrogenases b-type cytochrome subunit signature 1.
 CONSENSUS: R-[LIVMFYW]-x-H-W-[LIVM]-x(2)-[LIVMF]-[STAC]-[LIVM]-x(2)-L-x-[LIVM]-T-G.
- 15 NAME: Nickel-dependent hydrogenases b-type cytochrome subunit signature 2.
 CONSENSUS: [RH]-[STA]-[LIVMFYW]-H-[RH]-[LIVM]-x(2)-W-x-[LIVMF]-x(2)-F-x(3)-H.
- 20 NAME: Succinate dehydrogenase cytochrome b subunit signature 1.
 CONSENSUS: R-P-[LIVMT]-x(3)-[LIVM]-x(6)-[LIVMWPK]-x(4)-S-x(2)-H-R-x-[ST].
- 25 NAME: Succinate dehydrogenase cytochrome b subunit signature 2.
 CONSENSUS: H-x(3)-[GA]-[LIVMT]-R-[HF]-[LIVMF]-x-[FYWM]-D-x-[GVA].
- 30 NAME: Thioredoxin family active site.
 CONSENSUS: [LIVMF]-[LIVMSTA]-x-[LIVMFYC]-[FYWSTHE]-x(2)-[FYWGTN]-C-[GATPLVE]-
 CONSENSUS: [PHYWSTA]-C-x(6)-[LIVMFYWT].
- 35 NAME: Glutaredoxin active site.
 CONSENSUS: [LIVD]-[FYSA]-x(4)-C-[PV]-[FYW]-C-x(2)-[TAV]-x(2,3)-[LIV].
- 40 NAME: Type-1 copper (blue) proteins signature.
 CONSENSUS: [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ].
- 45 NAME: 2Fe-2S ferredoxins, iron-sulfur binding region signature.
 CONSENSUS: C-[C]-[C]-[GA]-[C]-C-[GAST]-[CPDEKRHFYW]-C.
- NAME: Adrenodoxin family, iron-sulfur binding region signature.
 CONSENSUS: C-x(2)-[STAQ]-x-[STAMV]-C-[STA]-T-C-[HR].
- 50 NAME: 4Fe-4S ferredoxins, iron-sulfur binding region signature.
 CONSENSUS: C-x(2)-C-x(2)-C-x(3)-C-[PEG].
- 55 NAME: High potential iron-sulfur proteins signature.
 CONSENSUS: C-x(6,9)-[LIVM]-x(3)-G-[YW]-C-x(2)-[FYW].
- NAME: Rieske iron-sulfur protein signature 1.
 CONSENSUS: C-[TK]-H-L-G-C-[LIVT].

- NAME: Rieske iron-sulfur protein signature 2.
 CONSENSUS: C-P-C-H-x-[GSA].
- 5 NAME: Flavodoxin signature.
 CONSENSUS: [LIV]-[LIVFY]-[FY]-x-[ST]-x(2)-[AGC]-x-T-x(3)-A-x(2)-[LIV].
- 10 NAME: Rubredoxin signature.
 CONSENSUS: [LIVM]-x(3)-W-x-C-P-x-C-[AGD].
- NAME: Electron transfer flavoprotein alpha-subunit signature.
 CONSENSUS: [LI]-Y-[LIVM]-[AT]-x-G-[IV]-[SD]-G-x-[IV]-Q-H-x(2)-G-x(6)-[IV]-x-A-
 15 CONSENSUS: [IV]-N.
- NAME: Electron transfer flavoprotein beta-subunit signature.
 CONSENSUS: [IVA]-x-[KR]-x(2)-[DE]-[GD]-[GDE]-x(1,2)-[EQ]-x-[LIV]-x(4)-P-x-[LIVM](2)-
 20 CONSENSUS: [TAC].
- NAME: Vertebrate metallothioneins signature.
 CONSENSUS: C-x-C-[GSTAP]-x(2)-C-x-C-x(2)-C-x-C-x(2)-C-x-K.
 25
- NAME: Ferritin iron-binding regions signature 1.
 CONSENSUS: E-x-[KR]-E-x(2)-E-[KR]-[LF]-[LIVMA]-x(2)-Q-N-x-R-x-G-R.
- 30 NAME: Ferritin iron-binding regions signature 2.
 CONSENSUS: D-x(2)-[LIVMF]-[STAC]-[DH]-F-[LI]-[EN]-x(2)-[FY]-L-x(6)-[LIVM]-[KN].
- NAME: Bacterioferritin signature.
 35 CONSENSUS: <M-x-G-x(3)-V-[LIV]-x(2)-[LM]-x(3)-L-x(3)-L.
- NAME: Transferrins signature 1.
 CONSENSUS: Y-x(D,1)-[VAS]-V-[IVAC]-[IVA]-[IVA]-[RKH]-[RKS]-[GDENSA].
 40
- NAME: Transferrins signature 2.
 CONSENSUS: Y-x-G-A-[FL]-[KRHNQ]-C-L-x(3,4)-G-[DENQ]-V-[GA]-[FYW].
- 45 NAME: Transferrins signature 3.
 CONSENSUS: [DENQ]-[YF]-x-[LY]-L-C-x-[DN]-x(5,8)-[LIV]-x(4,5)-C-x(2)-A-x(4)-[HQR]-x-
 CONSENSUS: [LIVMFYW]-[LIVM].
- 50 NAME: Globins profile.
- NAME: Protozoan/cyanobacterial globins signature.
 CONSENSUS: F-[LF]-x(5)-G-[PA]-x(4)-G-[KRA]-x-[LIVM]-x(3)-H.
- 55 NAME: Plant hemoglobins signature.
 CONSENSUS: [SN]-P-x-L-x(2)-H-A-x(3)-F.
- NAME: Hemerythrins signature.

CONSENSUS: W-L-x-[ENQ]-H-I-x(3)-D-F.

NAME: Arthropod hemocyanins / insect LSPs signature 1.
CONSENSUS: Y-[FYW]-x-E-D-[LIVM]-x(2)-N-x(6)-H-x(3)-P.

NAME: Arthropod hemocyanins / insect LSPs signature 2.
CONSENSUS: T-x(2)-R-D-P-x-[FY]-[FYW].

NAME: Heavy-metal-associated domain.
CONSENSUS: [LIVN]-x(2)-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-
[LIVFG]-x(3)-[LIV]-x(9,11)-
CONSENSUS: [IVA]-x-[LVFYS].

NAME: ABC transporters family signature.
CONSENSUS: [LIVMFYC]-[SA]-[SAPGLVFYKQH]-G-[DENQMW]-
[KRQASPCLIMFW]-[KRNQSTAVM]-
CONSENSUS: [KRACLVM]-[LIVMFYPAN]-[PHY]-[LIVMFW]-[SAGCLIVP]-
[FYWHP]-[KRHP]-
CONSENSUS: [LIVMFYWSTA].

NAME: Binding-protein-dependent transport systems inner
membrane comp. sign.
CONSENSUS: [LIVMFY]-x(8)-[EQR]-[STAGV]-[STAG]-x(3)-G-
[LIVMFYSTAC]-x(5)-[LIVMFYSTA]-
CONSENSUS: x(4)-[LIVMFY]-[PKR].

NAME: ABC-2 type transport system integral membrane proteins
signature.
CONSENSUS: [LIMST]-x(2)-[LIMW]-x(2)-[LIMCA]-[GSTC]-x-[GSAIV]-
x(6)-[LIMGA]-[PGSNQ]-
CONSENSUS: x(9,12)-P-[LIMFT]-x-[HRSY]-x(5)-[RQ].

NAME: Bacterial extracellular solute-binding proteins,
family 1 signature.
CONSENSUS: [GAP]-[LIVMFA]-[STAVDN]-x(4)-[GSAV]-[LIVMFY](2)-Y-
[ND]-x(3)-[LIVMF]-x-
CONSENSUS: [KNDE].

NAME: Bacterial extracellular solute-binding proteins,
family 3 signature.
CONSENSUS: G-[FYIL]-[DED]-[LIVMT]-[DED]-[LIVMF]-x(3)-[LIVMA]-
[VAGC]-x(2)-[LIVMAGN].

NAME: Bacterial extracellular solute-binding proteins,
family 5 signature.
CONSENSUS: [AG]-x(6,7)-[DNEG]-x(2)-[STAVE]-[LIVMFYW]-x-
[LIVMFY]-x-[LIVM]-[KR]-
CONSENSUS: [KRHDE]-[GDN]-[LIVMA]-[KNGSP]-[FW].

NAME: Serum albumin family signature.
CONSENSUS: [FY]-x(6)-C-C-x(7)-C-[LFY]-x(6)-[LIVMFYW].

NAME: Transthyretin signature 1.
CONSENSUS: S-K-C-P-L-M-V-K-V-L-D-[AS]-V-R-G.

NAME: Transthyretin signature 2.
CONSENSUS: S-P-[FY]-S-[FY]-S-T-T-A-[LIVM]-V-[ST]-x-P.

- NAME: Avidin / Streptavidin family signature.
 CONSENSUS: [DEN]-x(2)-[KR]-[STA]-x(2)-V-G-x-[DN]-x-[FW]-T-[KR].
- 5 NAME: Eukaryotic cobalamin-binding proteins signature.
 CONSENSUS: [SN]-V-D-T-[GA]-A-[LIVM]-A-x-L-A-[LIVMF]-T-C.
- NAME: Lipocalin signature.
 CONSENSUS: [DENG]-x-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-
 10 [CP]-G-[C]-W-[FYWLRH]-x-
 CONSENSUS: [LIVMTA].
- NAME: Cytosolic fatty-acid binding proteins signature.
 CONSENSUS: [GSAIVK]-x-[FYW]-x-[LIVMF]-x(4)-[NHG]-[FY]-[DE]-x-
 15 [LIVMFY]-[LIVM]-x(2)-
 CONSENSUS: [LIVMAKR].
- NAME: Acyl-CoA-binding protein signature.
 CONSENSUS: P-[STA]-x-[DEN]-x-[LIVMF]-x(2)-[LIVMFY]-Y-[GSTA]-
 20 x-[FY]-K-Q-[STA](2)-x-G.
- NAME: LBP / BPI / CETP family signature.
 CONSENSUS: [PA]-[GA]-[LIVMC]-x(2)-R-[IV]-[ST]-x(3)-L-x(5)-
 [EQ]-x(4)-[LIVM]-[EQK]-
 25 CONSENSUS: x(8)-P.
- NAME: Phosphatidylethanolamine-binding protein family
 signature.
 CONSENSUS: [FY]-x-[LIVMF](3)-x-[DC]-P-D-x-P-[SN]-x(10)-H.
 30
- NAME: Plant lipid transfer proteins signature.
 CONSENSUS: [LIVM]-[PA]-x(2)-C-x-[LIVM]-x-[LIVM]-x-[LIVMFY]-x-
 [LIVM]-[ST]-x(3)-
 35 CONSENSUS: [DN]-C-x(2)-[LIVM].
- NAME: Uteroglobin family signature 1.
 CONSENSUS: [GA]-x(3)-I-C-P-x-[LIVMF]-x(3)-[LIVM]-[DE]-x-
 [LIVMF](2).
- 40 NAME: Uteroglobin family signature 2.
 CONSENSUS: [DEQ]-x(4)-[SN]-x(5)-[DEQ]-x-I-x(2)-S-[PSE]-[LS]-
 C.
- NAME: Mitochondrial energy transfer proteins signature.
 45 CONSENSUS: P-x-[DE]-x-[LIVAT]-[RK]-x-[LRH]-[LIVMFY]-[QMAIGV].
- NAME: Sugar transport proteins signature 1.
 CONSENSUS: [LIVMSTAG]-[LIVMFSAG]-x(2)-[LIVMSA]-[DE]-x-
 [LIVMFYWA]-G-R-[RK]-x(4,6)-
 50 CONSENSUS: [GSTA].
- NAME: Sugar transport proteins signature 2.
 CONSENSUS: [LIVMF]-x-G-[LIVMFA]-x(2)-G-x(8)-[LIFY]-x(2)-[EQ]-
 x(6)-[RK].
 55
- NAME: LacY family proton/sugar symporters signature 1.
 CONSENSUS: G-[LIVM](2)-x-D-[RK]-L-G-L-[RK](2)-x-[LIVM](2)-W.

- NAME: LacY family proton/sugar symporters signature 2.
 5 CONSENSUS: P-x-[LIVMF](2)-N-R-[LIVM]-G-x-K-N-[STA]-[LIVM](3).
- NAME: PTR2 family proton/oligopeptide symporters signature 1.
 CONSENSUS: [GA]-[GAS]-[LIVMFYWA]-[LIVM]-[GAS]-D-x-[LIVMFYWT]-
 [LIVMFYW]-G-x(3)-[TAV]-
 CONSENSUS: [IV]-x(3)-[GSTAV]-x-[LIVMF]-x(3)-[GA].
- 10 NAME: PTR2 family proton/oligopeptide symporters signature 2.
 CONSENSUS: [FYT]-x(2)-[LMFY]-[FYV]-[LIVMFYWA]-x-[IVG]-N-
 [LIVMAG]-G-[GSA]-[LIMF].
- 15 NAME: Amiloride-sensitive sodium channels signature.
 CONSENSUS: Y-x(2)-[EQTF]-x-C-x(2)-[GSTDL]-C-x-[QT]-x(2)-
 [LIVMT]-[LIVMS]-x(2)-C-x-C.
- 20 NAME: Sodium:alanine symporter family signature.
 CONSENSUS: G-G-x-[GA](2)-[LIVM]-F-W-M-W-[LIVM]-x-[STAV]-
 [LIVMFA](2)-G.
- 25 NAME: Sodium:dicarboxylate symporter family signature 1.
 CONSENSUS: P-x(D,L)-G-[DE]-x-[LIVMF](2)-x-[LIVM](2)-[KREQ]-
 [LIVM](3)-x-P.
- NAME: Sodium:dicarboxylate symporter family signature 2.
 CONSENSUS: P-x-G-x-[STA]-x-[NT]-[LIVMC]-D-G-[STAN]-x-[LIVM]-
 [FY]-x(2)-[LIVM]-x(2)-
 30 CONSENSUS: [LIVM]-[FY]-[LI]-[SA]-Q.
- NAME: Sodium:galactoside symporter family signature.
 CONSENSUS: D-x(3)-G-x(3)-[DN]-x(6,8)-G-[KH]-F-[KR]-P-[FYW]-
 [LIVM](2)-x-[GSTA](2).
- 35 NAME: Sodium:neurotransmitter symporter family signature 1.
 CONSENSUS: W-R-F-[GP]-Y-x(4)-N-G-G-G-x-[FY].
- NAME: Sodium:neurotransmitter symporter family signature 2.
 40 CONSENSUS: Y-[LIVMFY]-x(2)-[SC]-[LIVMFY]-[STQ]-x(2)-L-P-W-
 x(2)-C-x(4)-N-[GST].
- NAME: Sodium:solute symporter family signature 1.
 CONSENSUS: [GS]-x(2)-[LIY]-x(3)-[LIVMFYWSTAG](10)-[LIY]-
 45 [TAV]-x(2)-G-G-[LMF]-x-
 CONSENSUS: [SAP].
- NAME: Sodium:solute symporter family signature 2.
 CONSENSUS: [GAST]-[LIVM]-x(3)-[KR]-x(4)-G-A-x(2)-[GAS]-
 50 [LIVMGS]-[LIVMW]-[LIVMGAT]-G-
 CONSENSUS: x-[LIVMG].
- NAME: Sodium:sulfate symporter family signature.
 CONSENSUS: [STACP]-S-x(2)-F-x(2)-P-[LIVM]-[GSA]-x(3)-N-x-
 55 [LIVM]-V.
- NAME: glpT family of transporters signature.
 CONSENSUS: R-G-x(5)-W-N-x(2)-H-N-x-G-G.

- NAME: Ammonium transporters signature.
 CONSENSUS: D-[FYWS]-A-G-[GSC]-x(2)-[IV]-x(3)-[SAG](2)-x(2)-
 [SAG]-[LIVMF]-x(3)-
 5 CONSENSUS: [LIVMFYWA](2)-x-[GK]-x-R.
- NAME: BCCT family of transporters signature.
 CONSENSUS: [GSDN]-W-T-[LIVM]-x-[FY]-W-x-W-W.
- 10 NAME: Flagellar motor protein motA family signature.
 CONSENSUS: A-[LMF]-x-[GAT]-T-[LIVF]-x-G-x-[LIVMF]-x(?) -P.
- NAME: Formate and nitrite transporters signature 1.
 CONSENSUS: [LIVMA]-[LIVMY]-x-G-[GSTA]-[DES]-L-[FI]-[TN]-[GS].
 15
- NAME: Formate and nitrite transporters signature 2.
 CONSENSUS: [GA]-x(2)-[CA]-N-[LIVMFYW](2)-V-C-[LV]-A.
- NAME: Prokaryotic sulfate-binding proteins signature 1.
 CONSENSUS: K-x-[NQEK]-[GT]-G-[DQ]-x-[LIVM]-x(3)-Q-S.
 20
- NAME: Prokaryotic sulfate-binding proteins signature 2.
 CONSENSUS: N-P-K-[ST]-S-G-x-A-R.
- 25 NAME: Sulfate transporters signature.
 CONSENSUS: P-x-Y-[GS]-L-Y-[STAG](2)-x(4)-[LIVMFY](3)-x(3)-
 [GSTA](2)-S-[KR].
- NAME: Amino acid permeases signature.
 CONSENSUS: [STAGC]-G-[PAG]-x(2,3)-[LIVMFYWA](2)-x-[LIVMFYW]-
 x-[LIVMFYSTAGC](2)-
 CONSENSUS: [STAGC]-x(3)-[LIVMFYW]-x-[LIVMST]-x(3)-[LIMCTA]-
 [GA]-E-x(5)-[PSAL].
 30
- NAME: Aromatic amino acids permeases signature.
 CONSENSUS: I-G-[GA]-G-M-[LF]-[SA]-x-P-x(3)-[SA]-G-x(2)-F.
 35
- NAME: Xanthine/uracil permeases family signature.
 CONSENSUS: [LIVM]-P-x-[PASIF]-V-[LIVM]-G-G-x(4)-[LIVM]-[FY]-
 [GSA]-x-[LIVM]-x(3)-G.
 40
- NAME: Anion exchangers family signature 1.
 CONSENSUS: F-G-G-[LIVM](2)-[KR]-D-[LIVM]-[RK]-R-R-Y.
- 45 NAME: Anion exchangers family signature 2.
 CONSENSUS: [FI]-L-I-S-L-I-F-I-Y-E-T-F-x-K-L.
- NAME: MIP family signature.
 CONSENSUS: [HNQA]-x-N-P-[STA]-[LIVMF]-[ST]-[LIVMF]-[GSTAFY].
 50
- NAME: General diffusion Gram-negative porins signature.
 CONSENSUS: [LIVMFY]-x(2)-G-x(2)-Y-x-F-x-K-x(2)-[SN]-[STAV]-
 [LIVMFYW]-V.
- 55 NAME: OmpA-like domain.
 CONSENSUS: [LIVMA]-x-[GT]-x-[TA]-[DA]-x(2)-[DG]-[GSTP]-x(2)-
 [LFYDE]-[NQS]-x(2)-

- CONSENSUS: [LI]-[SG]-[QE]-[KRQE]-R-A-x(2)-[LV]-x(3)-[LIVMF]-
 x(4,5)-[LIVM]-x(4)-
 CONSENSUS: [LIVM]-x(3)-[SG]-x-G.
- 5 NAME: Eukaryotic mitochondrial porin signature.
 CONSENSUS: [YH]-x(2)-D-[SPA]-x-[STA]-x(3)-[TAG]-[KR]-[LIVMF]-
 [DNSTA]-[DNS]-x(4)-
 CONSENSUS: [GSTAN]-[LIVMA]-x-[LIVMY].
- 10 NAME: Insulin-like growth factor binding proteins signature.
 CONSENSUS: G-C-[GS]-C-C-x(2)-C-A-x(6)-C.
- NAME: GPR1/FUN34/yaaH family signature.
 CONSENSUS: N-P-[AV]-P-[LF]-G-L-x-[GSA]-F.
- 15 NAME: GNS1/SUR4 family signature.
 CONSENSUS: L-x-F-L-H-x-Y-H-H.
- 20 NAME: 43 Kd postsynaptic protein signature.
 CONSENSUS: G-Q-D-Q-T-K-Q-Q-I.
- NAME: Actins signature 1.
 CONSENSUS: [FY]-[LIV]-G-[DE]-E-A-Q-x-[RKQ](2)-G.
- 25 NAME: Actins signature 2.
 CONSENSUS: W-[IV]-[STA]-[RK]-x-[DE]-Y-[DNE]-[DE].
- NAME: Actins and actin-related proteins signature.
 CONSENSUS: [LM]-[LIVM]-T-E-[GAPQ]-x-[LIVMFYWHQ]-N-[PSTAQ]-
 30 x(2)-N-[KR].
- NAME: Annexins repeated domain signature.
 CONSENSUS: [TG]-[STV]-x(8)-[LIVMF]-x(2)-R-x(3)-[DEQNH]-x(7)-
 [IFY]-x(7)-[LIVMF]-
 35 CONSENSUS: x(3)-[LIVMF]-x(11)-[LIVMFA]-x(2)-[LIVMF].
- NAME: Caveolins signature.
 CONSENSUS: F-E-D-V-I-A-E-P.
- 40 NAME: Clathrin light chain signature 1.
 CONSENSUS: F-L-A-Q-Q-E-S.
- NAME: Clathrin light chain signature 2.
 CONSENSUS: [KR]-D-x-S-[KR]-[LIVM]-[KR]-x-[LIVM](3)-x-L-K.
- 45 NAME: Clusterin signature 1.
 CONSENSUS: C-K-P-C-L-K-x-T-C.
- NAME: Clusterin signature 2.
 50 CONSENSUS: C-L-[RK]-M-[RK]-x-[EQ]-C-[ED]-K-C.
- NAME: Connexins signature 1.
 CONSENSUS: C-[DN]-T-x-Q-P-G-C-x(2)-V-C-Y-D.
- 55 NAME: Connexins signature 2.
 CONSENSUS: C-x(3,4)-P-C-x(3)-[LIVM]-[DEN]-C-[FY]-[LIVM]-[SA]-
 [KR]-P.

- NAME: Crystallins beta and gamma 'Greek key' motif
signature.
CONSENSUS: [LIVMFYWAA]-x-[DEHRKSTP]-[FY]-[DEQHKY]-x(3)-[FY]-x-
G-x(4)-[LIVMFCST].
- 5 NAME: Dynamin family signature.
CONSENSUS: L-P-[RK]-G-[STN]-[GN]-[LIVM]-V-T-R.
- 10 NAME: Dynein light chain type 1 signature.
CONSENSUS: H-x-I-x-G-[KR]-x-F-[GA]-S-x-V-[ST]-[HY]-E.
- NAME: FtsZ protein signature 1.
CONSENSUS: N-[ST]-D-x-Q-x-L-x(16,18)-G-x-G-[ATV]-G-[GSAN]-x-
P-x(2)-G.
- 15 NAME: FtsZ protein signature 2.
CONSENSUS: [DNHKR]-[LIVMF]-x-[LIVMF](2)-[VSTAC]-[STAC]-G-x-G-
[GK]-G-T-G-[ST]-G-
CONSENSUS: [GSAR]-[STA]-P-[LIVMFT]-[LIVMF]-[SGAV].
- 20 NAME: Fungal hydrophobins signature.
CONSENSUS: [GN]-[DNQPSA]-x-C-[GSTANK]-[GSTADNQ]-[STNQI]-
[PTIV]-x-C-C-[DENQKPST].
- 25 NAME: Intermediate filaments signature.
CONSENSUS: [IV]-x-[TACI]-Y-[RKH]-x-[LM]-L-[DE].
- NAME: Involucrin signature.
CONSENSUS: <M-S-[QH]-Q-x-T-[LV]-P-V-T-[LV].
- 30 NAME: Kinesin motor domain signature.
CONSENSUS: [GSA]-[KRHPSTQVM]-[LIVMF]-x-[LIVMF]-[IVC]-D-L-
[AH]-G-[SAN]-E.
- 35 NAME: Kinesin motor domain profile.
- NAME: Kinesin light chain repeat.
CONSENSUS: [DEQR]-A-L-x(3)-[GEQ]-x(3)-G-x-[DNS]-x-P-x-V-A-
x(3)-N-x-L-[AS]-
40 CONSENSUS: x(5)-[QR]-x-[KR]-[FY]-x(2)-[AV]-x(4)-[HKNQ].
- NAME: Myelin basic protein signature.
CONSENSUS: V-V-H-F-F-K-N.
- 45 NAME: Myelin PD protein signature.
CONSENSUS: S-[KR]-S-x-K-[AG]-x-[SA]-E-K-K-[STA]-K.
- NAME: Myelin proteolipid protein signature 1.
CONSENSUS: G-[MV]-A-L-F-C-G-C-G-H.
- 50 NAME: Myelin proteolipid protein signature 2.
CONSENSUS: C-x-[ST]-x-[DE]-x(3)-[ST]-[FY]-x-L-[FY]-I-x(4)-G-
A.
- 55 NAME: Neuromodulin (GAP-43) signature 1.
CONSENSUS: <M-L-C-C-[LIVM]-R-R.
- NAME: Neuromodulin (GAP-43) signature 2.

- CONSENSUS: S-F-R-G-H-I-x-R-K-K-[LIVM].
- NAME: Osteopontin signature.
 5 CONSENSUS: [EKQ]-x-[TA]-x(2)-[GA]-S-S-E-E-K.
- NAME: Peripherin / rom-1 signature.
 CONSENSUS: D-[GS]-V-P-F-[ST]-C-C-N-P-x-S-P-R-P-C.
- NAME: Profilin signature.
 10 CONSENSUS: <x(0,1)-[STA]-x(0,1)-W-[DENQH]-x-[YI]-x-[DEQ].
- NAME: Surfactant associated polypeptide SP-C palmitoylation sites.
 CONSENSUS: I-P-C-C-P-V.
 15
- NAME: Synapsins signature 1.
 CONSENSUS: L-R-R-R-L-S-D-S.
- NAME: Synapsins signature 2.
 20 CONSENSUS: G-H-A-H-S-G-M-G-K-V-K.
- NAME: Synaptobrevin signature.
 CONSENSUS: N-[LIVM]-[DENS]-[KL]-V-x-[DEQ]-R-x(2)-[KR]-[LIVM]-[STDE]-x-[LIVM]-x-[DE]-
 25 CONSENSUS: [KR]-[TA]-[DE].
- NAME: Synaptophysin / synaptoporin signature.
 CONSENSUS: L-S-V-[DE]-C-x-N-K-T.
- NAME: Tropomyosins signature.
 30 CONSENSUS: L-K-E-A-E-x-R-A-E.
- NAME: Tubulin subunits alpha, beta, and gamma signature.
 CONSENSUS: [SAG]-G-G-T-G-[SA]-G.
 35
- NAME: Tubulin-beta mRNA autoregulation signal.
 CONSENSUS: <M-R-[DE]-[IL].
- NAME: Tau and MAP proteins tubulin-binding domain signature.
 40 CONSENSUS: G-S-x(2)-N-x(2)-H-x-[PA]-[AG]-G(2).
- NAME: Neuraxin and MAP1B proteins repeated region signature.
 CONSENSUS: [STAGDN]-Y-x-Y-E-x(2)-[DE]-[KR]-[STAGCI].
- NAME: F-actin capping protein alpha subunit signature 1.
 45 CONSENSUS: V-H-[FY](2)-E-D-G-N-V.
- NAME: F-actin capping protein alpha subunit signature 2.
 CONSENSUS: F-K-[AE]-L-R-R-x-L-P.
 50
- NAME: F-actin capping protein beta subunit signature.
 CONSENSUS: C-D-Y-N-R-D.
- NAME: Vinculin family talin-binding region signature.
 55 CONSENSUS: [KR]-x-[LIVMF]-x(3)-[LIVMA]-x(2)-[LIVM]-x(6)-R-Q-Q-E-L.
- NAME: Vinculin repeated domain signature.

CONSENSUS: [LIVM]-x-[QA]-A-x(2)-W-[IL]-x-[DN]-P.

NAME: Amyloidogenic glycoprotein extracellular domain signature.

5 CONSENSUS: G-[VT]-E-[FY]-V-C-C-P.

NAME: Amyloidogenic glycoprotein intracellular domain signature.

10 CONSENSUS: G-Y-E-N-P-T-Y-[KR].

NAME: Cadherins extracellular repeated domain signature.

CONSENSUS: [LIV]-x-[LIV]-x-D-x-N-D-[NH]-x-P.

NAME: Insect cuticle proteins signature.

15 CONSENSUS: G-x(7)-[DEN]-G-x(6)-Y-x-A-[DNG]-x(2,3)-G-[FY]-x-[AP].

NAME: Gas vesicles protein GVPa signature 1.

20 CONSENSUS: [LIVM]-x-[DE]-[LIVMFYT]-[LIVM]-[DE]-x-[LIVM](2)-[DKR](2)-G-x-[LIVM](2).

NAME: Gas vesicles protein GVPa signature 2.

CONSENSUS: R-[LIVA](3)-A-[GS]-[LIVMFY]-x-T-x(3)-Y-[AG].

25 NAME: Gas vesicles protein GVPc repeated domain signature.

CONSENSUS: F-L-x(2)-T-x(3)-R-x(3)-A-x(2)-Q-x(3)-L-x(2)-F.

NAME: Bacterial microcompartments proteins signature.

30 CONSENSUS: D-x(0,1)-M-x-K-[SAG](2)-x-[IV]-x-[LIVM]-[LIVMA]-[GCS]-x(4)-[GD]-[SGPD]-
CONSENSUS: [GA].

NAME: Flagella basal body rod proteins signature.

35 CONSENSUS: [GTARYQ]-x(9)-[LIVMYSTA](2)-[GSTA]-[STADEN]-N-[LIVM]-[SAN]-N-x-[SADNFR]-
CONSENSUS: [STV].

NAME: Flagella transport protein fliP family signature 1.

40 CONSENSUS: [PA]-A-[FY]-x-[LIVT]-[STH]-[EQ]-[LI]-x(2)-[GA]-F-[KREQ]-[IM]-G-[LIF].

NAME: Flagella transport protein fliP family signature 2.

CONSENSUS: P-[LIVMF]-K-[LIVMF](5)-x-[LIVMA]-[DNGS]-G-W.

45 NAME: Plant viruses icosahedral capsid proteins 'S' region signature.

CONSENSUS: [FYW]-x-[PSTA]-x(7)-G-x-[LIVM]-x-[LIVM]-x-[FYWI]-x(2)-D-x(5)-P.

50 NAME: Potexviruses and carlaviruses coat protein signature.

CONSENSUS: [RK]-[FYW]-A-[GAP]-F-D-x-F-x(2)-[LV]-x(3)-[GAST](2).

NAME: Neurotransmitter-gated ion-channels signature.

55 CONSENSUS: C-x-[LIVMFQ]-x-[LIVMF]-x(2)-[FY]-P-x-D-x(3)-C.

NAME: ATP P2X receptors signature.

CONSENSUS: G-G-x-[LIVM]-G-[LIVM]-x-[IV]-x-W-x-C-[DN]-L-D-x(5)-C-x-P-x-Y-x-F.

NAME: G-protein coupled receptors signature.

5 CONSENSUS: [GSTALIVMFYWC]-[GSTANCPDE]-[EDPKRH]-x(2)-[LIVMNQGA]-x(2)-[LIVMFT]-

CONSENSUS: [GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)-[LIVM].

10 NAME: G-protein coupled receptors family 2 signature 1.

CONSENSUS: C-x(3)-[FYWLIV]-D-x(3,4)-C-[FW]-x(2)-[STAGV]-x(8,9)-C-[PF].

NAME: G-protein coupled receptors family 2 signature 2.

15 CONSENSUS: Q-G-[LMFCA]-[LIVMFT]-[LIV]-x-[LIVFST]-[LIF]-[VFYH]-C-[LFY]-x-N-x(2)-V.

NAME: G-protein coupled receptors family 3 signature 1.

20 CONSENSUS: [LV]-x-N-[LIVM](2)-x-L-F-x-I-[PA]-Q-[LIVM]-[STA]-x-[STA](3)-[STAN].

NAME: G-protein coupled receptors family 3 signature 2.

25 CONSENSUS: C-C-[FYW]-x-C-x(2)-C-x(4)-[FYW]-x(2,4)-[DN]-x(2)-[STAH]-C-x(2)-C.

NAME: G-protein coupled receptors family 3 signature 3.

CONSENSUS: F-N-E-[STA]-K-x-I-[STAG]-F-[ST]-M.

NAME: Visual pigments (opsins) retinal binding site.

30 CONSENSUS: [LIVMWAC]-[PGAC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-[STACP]-x(2)-[DENF]-

CONSENSUS: [AP]-x(2)-[IY].

NAME: Bacterial rhodopsins signature 1.

35 CONSENSUS: R-Y-x-[DT]-W-x-[LIVMF]-[ST]-T-P-[LIVM](3).

NAME: Bacterial rhodopsins retinal binding site.

40 CONSENSUS: [FYIV]-x-[FYVG]-[LIVM]-D-[LIVMF]-x-[STA]-K-x(2)-[FY].

NAME: Receptor tyrosine kinase class II signature.

CONSENSUS: [DN]-[LIV]-Y-x(3)-Y-Y-R.

NAME: Receptor tyrosine kinase class III signature.

45 CONSENSUS: G-x-H-x-N-[LIVM]-V-N-L-L-G-A-C-T.

NAME: Receptor tyrosine kinase class V signature 1.

50 CONSENSUS: F-x-[DN]-x-[GAW]-[GA]-C-[LIVM]-[SA]-[LIVM](2)-[SA]-[LV]-[KRRH]-[LIVA]-

CONSENSUS: x(3)-[KR]-C-[PSAW].

NAME: Receptor tyrosine kinase class V signature 2.

55 CONSENSUS: C-x(2)-[DE]-G-[DEQ]-W-x(2,3)-[PAQ]-[LIVMT]-[GT]-x-C-x-C-x(2)-G-[HFY]-

CONSENSUS: [EQ].

NAME: Growth factor and cytokines receptors family signature 1.

CONSENSUS: C-[LVFYR]-x(7,8)-[STIVDN]-C-x-W.

NAME: Growth factor and cytokines receptors family signature 2.

5 CONSENSUS: [STGL]-x-W-[SG]-x-W-S.

NAME: TNFR/NGFR family cysteine-rich region signature.

CONSENSUS: C-x(4,6)-[FYH]-x(5,10)-C-x(0,2)-C-x(2,3)-C-x(7,11)-C-x(4,6)-[DNEQSKP]-

10 CONSENSUS: x(2)-C.

NAME: TNFR/NGFR family cysteine-rich region domain.

NAME: Integrins alpha chain signature.

15 CONSENSUS: [FYWS]-[RK]-x-G-F-F-x-R.

NAME: Integrins beta chain cysteine-rich domain signature.

CONSENSUS: C-x-[GNQ]-x(1,3)-G-x-C-x-C-x(2)-C-x-C.

20 NAME: Natriuretic peptides receptors signature.

CONSENSUS: G-P-x-C-x-Y-x-A-A-x-V-x-R-x(3)-H-W.

NAME: Photosynthetic reaction center proteins signature.

25 CONSENSUS: [NH]-x(4)-P-x-H-x(2)-[SAG]-x(11)-[SAGC]-x-H-[SAG](2).

NAME: Antenna complexes alpha subunits signature.

CONSENSUS: [LIVFAG]-x-[GASV]-[LIVFA]-x-[IV]-H-x(3)-[LIVM]-[GSTAE]-[STANH]-x(1,3)-

30 CONSENSUS: [STN]-W-[LIVMFYW].

NAME: Antenna complexes beta subunits signature.

CONSENSUS: [EQ]-x(4)-H-x(5)-[GSTA]-x(3)-[FY]-x(3)-[AG]-x(2)-[AV]-H-x(7)-P.

35

NAME: Photosystem I psaA and psaB proteins signature.

CONSENSUS: C-D-G-P-G-R-G-G-T-C.

NAME: Photosystem I psaG and psaK proteins signature.

40 CONSENSUS: G-F-x-[LIVM]-x-[DEA]-x(2)-[GA]-x-[GTA]-[SA]-x-G-H-x-[LIVM]-[GA].

NAME: Phytochrome chromophore attachment site signature.

45 CONSENSUS: [RGS]-[GSA]-[PV]-H-x-C-H-x(2)-Y.

NAME: Phytochrome chromophore attachment site domain profile.

NAME: Speract receptor repeated domain signature.

50 CONSENSUS: G-x(5)-G-x(2)-E-x(6)-W-G-x(2)-C-x(3)-[FYW]-x(8)-C-x(3)-G.

NAME: TonB-dependent receptor proteins signature 1.

55 CONSENSUS: <x(10,115)-[DENF]-[ST]-[LIVMF]-[LIVSTEQ]-V-x-[AGP]-[STANEQPK].

NAME: TonB-dependent receptor proteins signature 2.

CONSENSUS: [LYGSTANE]-x(3)-[EGSTAENQ]-x-[PGE]-R-x-[LIVFYWA]-x-
[LIVMFTA]-[ESTAGNQ]-
CONSENSUS: [LIVMFYGT]-x-[LIVMFYWGTAQ]-x-F>.

- 5 NAME: Transmembrane 4 family signature.
CONSENSUS: G-x(3)-[LIVMF]-x(2)-[GSA]-[LIVMF](2)-G-C-x-[GA]-
[STA]-x(2)-[EG]-x(2)-
CONSENSUS: [CWN]-[LIVM](2).
- 10 NAME: Bacterial chemotaxis sensory transducers signature.
CONSENSUS: R-T-E-[EQ]-Q-x(2)-[SA]-[LIVM]-x-[EQ]-T-A-A-S-M-E-
Q-L-T-A-T-V.
- 15 NAME: ER lumen protein retaining receptor signature 1.
CONSENSUS: G-I-S-x-[KR]-x-Q-x-L-[FY]-x-[LIV](2)-F-x(2)-R-Y.
NAME: ER lumen protein retaining receptor signature 2.
CONSENSUS: L-E-[SA]-V-A-I-[LM]-P-Q-L.
- 20 NAME: Ephrins signature.
CONSENSUS: [KRQ]-[LF]-[CST]-x-K-[IF]-Q-x-[FY]-[ST]-[PA]-x(3)-
G-x-E-F-x(5)-[FY](2)-
CONSENSUS: x(2)-[SA].
- 25 NAME: Granulins signature.
CONSENSUS: C-x-D-x(2)-H-C-C-P-x(4)-C.
NAME: HBGF/FGF family signature.
CONSENSUS: G-x-L-x-[STAGP]-x(6,7)-[DE]-C-x-[FM]-x-E-x(6)-Y.
- 30 NAME: PTN/MK heparin-binding protein family signature 1.
CONSENSUS: S-[DE]-C-x-[DE]-W-x-W-x(2)-C-x-P-x-[SN]-x-D-C-G-
[LIVMA]-G-x-R-E-G.
- 35 NAME: PTN/MK heparin-binding protein family signature 2.
CONSENSUS: C-[KR]-[LIVM]-P-C-N-W-K-K-x-F-G-A-[DE]-C-K-Y-x-F-
[EQ]-x-W-G-x-C.
- 40 NAME: Nerve growth factor family signature.
CONSENSUS: G-C-[KR]-G-[LIV]-[DE]-x(3)-[YW]-x-S-x-C.
NAME: Platelet-derived growth factor (PDGF) family
signature.
CONSENSUS: P-[PS]-C-V-x(3)-R-C-[GSTA]-G-C-C.
- 45 NAME: Small cytokines (intercrine/chemokine) C-x-C subfamily
signature.
CONSENSUS: C-x-C-[LIVM]-x(5,6)-[LIVMFY]-x(2)-[RKSEQ]-x-
[LIVM]-x(2)-[LIVM]-x(5)-
50 CONSENSUS: [SAG]-x(2)-C-x(3)-[EQ]-[LIVM](2)-x(9,10)-C-L-[DN].
NAME: Small cytokines (intercrine/chemokine) C-C subfamily
signature.
CONSENSUS: C-C-[LIFYT]-x(5,6)-[LI]-x(4)-[LIVMF]-x(2)-[FYW]-
55 x(6,8)-C-x(3,4)-[SAG]-
CONSENSUS: [LIVM](2)-[FL]-x(8)-C-[STA].
NAME: TGF-beta family signature.

- CONSENSUS: [LIVM]-x(2)-P-x(2)-[FY]-x(4)-C-x-G-x-C.
- NAME: TNF family signature.
- 5 CONSENSUS: [LV]-x-[LIVM]-x(3)-G-[LIVMF]-Y-[LIVMFY](2)-x(2)-[QEKHL]-[LIVMGT]-x-
- CONSENSUS: [LIVMFY].
- NAME: TNF family profile.
- 10 NAME: Wnt-1 family signature.
- CONSENSUS: C-K-C-H-G-[LIVMT]-S-G-x-C.
- NAME: Interferon alpha, beta and delta family signature.
- 15 CONSENSUS: [FYH]-[FY]-x-[GNRC]-[LIVM]-x(2)-[FY]-L-x(7)-[CY]-A-W.
- NAME: Granulocyte-macrophage colony-stimulating factor signature.
- 20 CONSENSUS: C-P-[LP]-T-x-E-[ST]-x-C.
- NAME: Interleukin-1 signature.
- CONSENSUS: [FC]-x-S-[ASLV]-x(2)-P-x(2)-[FYLIV]-[LI]-[SCA]-T-x(7)-[LIVM].
- 25 NAME: Interleukin-2 signature.
- CONSENSUS: T-E-[LF]-x(2)-L-x-C-L-x(2)-E-L.
- NAME: Interleukins -4 and -13 signature.
- 30 CONSENSUS: L-x-E-[LIVM](2)-x(4,5)-[LIVM]-[TL]-x(5,7)-C-x(4)-[IVA]-x-[DNS]-[LIVMA].
- NAME: Interleukin-6 / G-CSF / MGF signature.
- CONSENSUS: C-x(9)-C-x(6)-G-L-x(2)-[FY]-x(3)-L.
- 35 NAME: Interleukin-7 and -9 signature.
- CONSENSUS: N-x-[LAP]-[SCT]-F-L-K-x-L-L.
- NAME: Interleukin-10 family signature.
- 40 CONSENSUS: [GS]-C-x(2)-[LV]-x(2)-[LIVM](2)-x-F-Y-L-x(2)-V.
- NAME: LIF / OSM family signature.
- CONSENSUS: [PST]-x(4)-F-[NQ]-x-K-x(3)-C-x-[LF]-L-x(2)-Y-[HK].
- 45 NAME: Macrophage migration inhibitory factor family signature.
- CONSENSUS: [DE]-P-C-A-x(3)-[LIVM]-x-S-I-G-x-[LIVM]-G.
- NAME: Adipokinetin hormone family signature.
- 50 CONSENSUS: Q-[LV]-[NT]-[FY]-[ST]-x(2)-W.
- NAME: Bombesin-like peptides family signature.
- CONSENSUS: W-A-x-G-[SH]-[LF]-M.
- 55 NAME: Calcitonin / CGRP / IAPP family signature.
- CONSENSUS: C-[SAGDN]-[STN]-x(0,1)-[SA]-T-C-[VMA]-x(3)-[LYF]-x(3)-[LYF].
- NAME: Corticotropin-releasing factor family signature.

- CONSENSUS: [PQ]-x-[LIVM]-S-[LIVM]-x(2)-[PST]-[LIVMF]-x-[LIVM]-L-R-x(2)-[LIVM].
- 5 NAME: Crustacean CHH/MIH/GIH neurohormones family signature.
 CONSENSUS: C-[DENK]-D-C-x-N-[LIV]-[FY]-R-x(7)-C-[KR]-x(2)-C.
- NAME: Erythropoietin / thrombopoietin signature.
 CONSENSUS: P-x(4)-C-D-x-R-[LIVM](2)-x-[KR]-x(14)-C.
- 10 NAME: Granins signature 1.
 CONSENSUS: [DE]-[SN]-L-[SAN]-x(2)-[DE]-x-E-L.
- NAME: Granins signature 2.
 15 CONSENSUS: C-[LIVM](2)-E-[LIVM](2)-S-[DN]-[STA]-L-x-K-x-S-x(3)-[LIVM]-[STA]-x-E-C.
- NAME: Galanin signature.
 CONSENSUS: G-W-T-L-N-S-A-G-Y-L-L-G-P-H.
- 20 NAME: Gastrin / cholecystokinin family signature.
 CONSENSUS: Y-x(D,L)-[GD]-[WH]-M-[DR]-F.
- NAME: Glucagon / GIP / secretin / VIP family signature.
 25 CONSENSUS: [YH]-[STAI VGD]-[DEQ]-[AGF]-[LIVMSTE]-[FYLR]-x-[DENSTAK]-[DENSTA]-
 CONSENSUS: [LIVMFY G]-x(9)-[KREQL]-[KRDENQL]-[LVFYWG]-[LIVQ].
- NAME: Glycoprotein hormones alpha chain signature 1.
 30 CONSENSUS: C-x-G-C-C-[FY]-S-R-A-[FY]-P-T-P.
- NAME: Glycoprotein hormones alpha chain signature 2.
 CONSENSUS: N-H-T-x-C-x-C-x-T-C-x(2)-H-K.
- NAME: Glycoprotein hormones beta chain signature 1.
 35 CONSENSUS: C-[STAGM]-G-[HFYL]-C-x-[ST].
- NAME: Glycoprotein hormones beta chain signature 2.
 CONSENSUS: [PA]-V-A-x(2)-C-x-C-x(2)-C-x(4)-[STD]-[DEY]-C-x(6,8)-[PGSTAVM]-x(2)-C.
- 40 NAME: Gonadotropin-releasing hormones signature.
 CONSENSUS: Q-H-[FYW]-S-x(4)-P-G.
- NAME: Insulin family signature.
 45 CONSENSUS: C-C-[P]-x(2)-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C.
- NAME: Natriuretic peptides signature.
 CONSENSUS: C-F-G-x(3)-D-R-I-x(3)-S-x(2)-G-C.
- 50 NAME: Neurohypophysial hormones signature.
 CONSENSUS: C-[LIFY](2)-x-N-[CS]-P-x-G.
- NAME: Neuromedin U signature.
 CONSENSUS: F-[LIVMF]-F-R-P-R-N.
- 55 NAME: Endogenous opioids neuropeptides precursors signature.
 CONSENSUS: C-x(3)-C-x(2)-C-x(2)-[KRH]-x(6,7)-[LIF]-[DN]-x(3)-C-x-[LIVM]-[EQ]-C-

- 5
 10
 15
 20
 25
 30
 35
 40
 45
 50
 55
- CONSENSUS: [EQ]-x(8)-W-x(2)-C.
 NAME: Pancreatic hormone family signature.
 CONSENSUS: [FY]-x(3)-[LIVM]-x(2)-Y-x(3)-[LIVMFY]-x-R-x-R-[YF].
 NAME: Parathyroid hormone family signature.
 CONSENSUS: V-S-E-x-Q-x(2)-H-x(2)-G.
 NAME: Pyrokinins signature.
 CONSENSUS: F-[GSTV]-P-R-L-[G>].
 NAME: Somatotropin, prolactin and related hormones signature 1.
 CONSENSUS: C-x-[ST]-x(2)-[LIVMFY]-x-[LIVMSTA]-P-x(5)-[TALIV]-x(7)-[LIVMFY]-x(6)-
 CONSENSUS: [LIVMFY]-x(2)-[STA]-W.
 NAME: Somatotropin, prolactin and related hormones signature 2.
 CONSENSUS: C-[LIVMFY]-x(2)-D-[LIVMFYSTA]-x(5)-[LIVMFY]-x(2)-[LIVMFYT]-x(2)-C.
 NAME: Tachykinin family signature.
 CONSENSUS: F-[IVFY]-G-[LM]-M-[G>].
 NAME: Thymosin beta-4 family signature.
 CONSENSUS: K-L-K-K-T-E-T-Q-E-K-N.
 NAME: Urotensin II signature.
 CONSENSUS: C-F-W-K-Y-C.
 NAME: Cecropin family signature.
 CONSENSUS: W-x(0,2)-[KDN]-x(2)-K-[KRE]-[LI]-E-[RKN].
 NAME: Mammalian defensins signature.
 CONSENSUS: C-x-C-x(3,5)-C-x(7)-G-x-C-x(9)-C-C.
 NAME: Arthropod defensins signature.
 CONSENSUS: C-x(2,3)-[HN]-C-x(3,4)-[GR]-x(2)-G-G-x-C-x(4,7)-C-x-C.
 NAME: Cathelicidins signature 1.
 CONSENSUS: Y-x-[ED]-x-V-x-[RQ]-A-[LIVMA]-[DQG]-x-[LIVMFY]-N-[EQ].
 NAME: Cathelicidins signature 2.
 CONSENSUS: F-x-[LIVM]-K-E-T-x-C-x(10)-C-x-F-[KR]-[KE].
 NAME: Endothelin family signature.
 CONSENSUS: C-x-C-x(4)-D-x(2)-C-x(2)-[FY]-C.
 NAME: Plant thionins signature.
 CONSENSUS: C-C-x(5)-R-x(2)-[FY]-x(2)-C.
 NAME: Gamma-thionins family signature.
 CONSENSUS: [KR]-x-C-x(3)-[SV]-x(2)-[FYWH]-x-[GF]-x-C-x(5)-C-x(3)-C.

- NAME: Snake toxins signature.
 CONSENSUS: G-C-x(1,3)-C-P-x(8,10)-C-C-x(2)-[P]DEN].
- 5 NAME: Myotoxins signature.
 CONSENSUS: K-x-C-H-x-K-x(2)-H-C-x(2)-K-x(3)-C-x(8)-K-x(2)-C-x(2)-[RK]-x-K-C-C-K-K.
- 10 NAME: Scorpion short toxins signature.
 CONSENSUS: C-x(3)-C-x(6,9)-[GAS]-K-C-[IMQT]-x(3)-C-x-C.
- NAME: Heat-stable enterotoxins signature.
 CONSENSUS: C-C-x(2)-C-C-x-P-A-C-x-G-C.
- 15 NAME: Aerolysin type toxins signature.
 CONSENSUS: [KT]-x(2)-N-W-x(2)-T-[DN]-T.
- NAME: Shiga/ricin ribosomal inactivating toxins active site signature.
 20 CONSENSUS: [LIVMA]-x-[LIVMSTA](2)-x-E-[SAGV]-[ESTAL]-R-[FY]-[RKNQS]-x-[LIVM]-[EQS]-
 CONSENSUS: ...x(2)-[LIVMF].
- NAME: Channel forming colicins signature.
 25 CONSENSUS: T-x(2)-W-x-P-[LIVMFY](3)-x(2)-E.
- NAME: Hok/gef family cell toxic proteins signature.
 CONSENSUS: [LIVMA](4)-C-[LIVMFA]-T-[LIVMA](2)-x(4)-[LIVM]-x-[RG]-x(2)-L-[CY].
- 30 NAME: Staphylococcal enterotoxin/Streptococcal pyrogenic exotoxin signature 1.
 CONSENSUS: Y-G-G-[LIV]-T-x(4)-N.
- 35 NAME: Staphylococcal enterotoxin/Streptococcal pyrogenic exotoxin signature 2.
 CONSENSUS: K-x(2)-[LIV]-x(4)-[LIV]-D-x(3)-R-x(2)-L-x(5)-[LIV]-Y.
- 40 NAME: Thiol-activated cytolysins signature.
 CONSENSUS: [RK]-E-C-T-G-L-x-W-E-W-W-[RK].
- NAME: Membrane attack complex components / perforin signature.
 45 CONSENSUS: Y-x(6)-[FY]-G-T-H-[FY].
- NAME: Pancreatic trypsin inhibitor (Kunitz) family signature.
 CONSENSUS: F-x(3)-G-C-x(6)-[FY]-x(5)-C.
- 50 NAME: Bowman-Birk serine protease inhibitors family signature.
 CONSENSUS: C-x(5,6)-[DENQKRHSTA]-C-[PASTDH]-[PASTDK]-[ASTDV]-C-[NDKS]-[DEKRHSTA]-C.
- 55 NAME: Kazal serine protease inhibitors family signature.
 CONSENSUS: C-x(7)-C-x(6)-Y-x(3)-C-x(2,3)-C.

- NAME: Soybean trypsin inhibitor (Kunitz) protease inhibitors family signature.
 5 CONSENSUS: [LIVM]-x-D-x-[EDNTY]-[DG]-[RKHDENQ]-x-[LIVM]-x(5)-Y-x-[LIVM].
- NAME: Serpins signature.
 10 CONSENSUS: [LIVMFY]-x-[LIVMFYAC]-[DNQ]-[RKHQSS]-[PST]-F-[LIVMFY]-[LIVMFYC]-x-
 CONSENSUS: [LIVMFAH].
- NAME: Potato inhibitor I family signature.
 15 CONSENSUS: [FYW]-P-[EQH]-[LIV](2)-G-x(2)-[STAGV]-x(2)-A.
- NAME: Squash family of serine protease inhibitors signature.
 15 CONSENSUS: C-P-x(5)-C-x(2)-D-x-D-C-x(3)-C-x-C.
- NAME: Streptomyces subtilisin-type inhibitors signature.
 CONSENSUS: C-x-P-x(2,3)-G-x-H-P-x(4)-A-C-[ATD]-x-L.
- 20 NAME: Cysteine proteases inhibitors signature.
 CONSENSUS: [GSTEQKRV]-Q-[LIVT]-[VAF]-[SAGQ]-G-x-[LIVMNK]-x(2)-[LIVMFY]-x-[LIVMFYA]-
 CONSENSUS: [DENQKRHSIV].
- 25 NAME: Tissue inhibitors of metalloproteinases signature.
 CONSENSUS: C-x-C-x-P-x-H-P-Q-x-A-F-C.
- NAME: Cereal trypsin/alpha-amylase inhibitors family signature.
 30 CONSENSUS: C-x(4)-[SAGD]-x(4)-[SPAL]-[LF]-x(2)-C-[RH]-x-[LIVMFY](2)-x(3,4)-C.
- NAME: Alpha-2-macroglobulin family thiolester region signature.
 35 CONSENSUS: [PG]-x-[GS]-C-[GA]-E-[EQ]-x-[LIVM].
- NAME: Disintegrins signature.
 CONSENSUS: C-x(2)-G-x-C-C-x-[NQRS]-C-x-[FM]-x(6)-C-[RK].
- 40 NAME: Lambdoid phages regulatory protein CIII signature.
 CONSENSUS: E-S-x-L-x-R-x(2)-[KR]-x-L-x(4)-[KR](2)-x(2)-[DE]-x-L.
- NAME: Chaperonins cpn60 signature.
 45 CONSENSUS: A-[AS]-x-[DEQ]-E-x(4)-G-G-[GA].
- NAME: Chaperonins cpn10 signature.
 CONSENSUS: [LIVMFY]-x-P-[ILT]-x-[DEN]-[KR]-[LIVMFA](3)-[KREQ]-x(8,9)-[SG]-x-
 50 CONSENSUS: [LIVMFY](3).
- NAME: Chaperonins TCP-1 signature 1.
 CONSENSUS: [RKEL]-[ST]-x-[LMFY]-G-P-x-[GSA]-x-x-K-[LIVMF](2).
- 55 NAME: Chaperonins TCP-1 signature 2.
 CONSENSUS: [LIVM]-[TS]-[NK]-D-[GA]-[AVNHK]-[TAV]-[LIVM](2)-x(2)-[LIVM]-x-[LIVM]-x-
 CONSENSUS: [SNH]-[PQH].

- NAME: Chaperonins TCP-1 signature 3.
 CONSENSUS: Q-[DEK]-x-x-[LIVMGTA]-[GA]-D-G-T.
- 5 NAME: Heat shock hsp20 proteins family profile.
 NAME: Heat shock hsp70 proteins family signature 1.
 CONSENSUS: [IV]-D-L-G-T-[ST]-x-[SC].
- 10 NAME: Heat shock hsp70 proteins family signature 2.
 CONSENSUS: [LIVMF]-[LIVMFY]-[DN]-[LIVMFS]-G-[GSH]-[GS]-[AST]-
 x(3)-[ST]-[LIVM]-
 CONSENSUS: [LIVMFC].
- 15 NAME: Heat shock hsp70 proteins family signature 3.
 CONSENSUS: [LIVMY]-x-[LIVMF]-x-G-G-x-[ST]-x-[LIVM]-P-x-
 [LIVM]-x-[DEQKRSTA].
- 20 NAME: Heat shock hsp90 proteins family signature.
 CONSENSUS: Y-x-[NQH]-K-[DE]-[IVA]-F-L-R-[ED].
 NAME: Chaperonins clpA/B signature 1.
 CONSENSUS: D-[AI]-[SGA]-N-[LIVMF](2)-K-[PT]-x-L-x(2)-G.
- 25 NAME: Chaperonins clpA/B signature 2.
 CONSENSUS: R-[LIVMFY]-D-x-S-E-[LIVMFY]-x-E-[KRQ]-x-[STA]-x-
 [STA]-[KR]-[LIVM]-x-G-
 CONSENSUS: [STA].
- 30 NAME: Nt-dnaJ domain signature.
 CONSENSUS: [FY]-x(2)-[LIVMA]-x(3)-[FYWHNT]-[DENQSA]-x-L-x-
 [DN]-x(3)-[KR]-x(2)-[FYI].
- NAME: dnaJ domain profile.
- 35 NAME: CXXCXGXG dnaJ domain signature.
 CONSENSUS: C-[DEGSTHKR]-x-C-x-G-x-[GK]-[AGSDM]-x(2)-[GSNKR]-
 x(4,6)-C-x(2,3)-C-x-G-x-G.
- 40 NAME: grpE protein signature.
 CONSENSUS: [FL]-[DN]-[PHEA]-x(2)-[HM]-x-A-[LIVMTN]-x(16,20)-
 G-[FY]-x(3)-[DEG]-x(2)-
 CONSENSUS: [LIVM]-[RI]-x-[SA]-x-V-x-[IV].
- 45 NAME: Bacterial type II secretion system protein C
 signature.
 CONSENSUS: P-x(6)-F-x(4)-L-x(3)-D-[LIVM]-A-[LIVM]-x-[LIVM]-N-
 x-[LIVM]-x-L.
- 50 NAME: Bacterial type II secretion system protein D
 signature.
 CONSENSUS: [GR]-[DEQKG]-[STVM]-[LIVMA](3)-[GA]-G-[LIVMFY]-
 x(11)-[LIVM]-P-
 CONSENSUS: [LIVMFYWG]-[LIVMF]-[GSAE]-x-[LIVM]-P-
 55 [LIVMFYW](2)-x(2)-[LV]-F.
- NAME: Bacterial type II secretion system protein E
 signature.

CONSENSUS: [LIVM]-R-x(2)-P-D-x-[LIVM](3)-G-E-[LIVM]-R-D.

NAME: Bacterial type II secretion system protein F signature.

5 CONSENSUS: [KRQ]-[LIVMA]-x(2)-[SAIV]-[LIVM]-x-[TY]-P-x(2)-[LIVM]-x(3)-[STAGV]-x(6)-

CONSENSUS: [LMY]-x(3)-[LIVMF](2)-P.

10 NAME: Bacterial type II secretion system protein N signature.

CONSENSUS: G-T-L-W-x-G-x(11)-L-x(4)-W.

NAME: Bacterial export FHIPEP family signature.

15 CONSENSUS: R-[LIVM]-[GSA]-E-V-[GSA]-A-R-F-[STV]-L-D-[GSA]-M-P-G-K-Q-M-[GSA]-I-D-

CONSENSUS: [GSA]-D.

NAME: Protein secA signatures.

20 CONSENSUS: [IV]-x-[IV]-[SA]-T-[NQ]-M-A-G-R-G-x-D-I-x-L.

NAME: Protein secY signature 1.

CONSENSUS: [GST]-[LIVMF](2)-x-[LIVM]-G-[LIVM]-x-P-[LIVMFY](2)-x-[AS]-[GSTQ]-

25 CONSENSUS: [LIVMFAT](3)-Q-[LIVMFA](2).

NAME: Protein secY signature 2.

CONSENSUS: [LIVMFYW](2)-x-[DE]-x-[LIVMF]-[STN]-x(2)-G-[LIVMF]-[GST]-[NST]-G-x-[GST]-

30 CONSENSUS: [LIVMF](3).

NAME: Protein secE/secE1-gamma signature.

CONSENSUS: [LIVMFY]-x(2)-[DENQGA]-x(4)-[LIVMTA]-x-[KRV]-x(2)-[KW]-P-x(3)-[SEQ]-x(7)-

35 CONSENSUS: [LIVT]-[LIVGA]-[LIVFGAST].

NAME: Gram-negative pili assembly chaperone signature.

CONSENSUS: [LIVMFY]-[APN]-x-[DNS]-[KREQ]-E-[STR]-[LIVMAR]-x-[FYWT]-x-[NC]-[LIVM]-

40 CONSENSUS: x(2)-[LIVM]-P-[PAS].

NAME: Fimbrial biogenesis outer membrane usher protein signature.

45 CONSENSUS: [VL]-[PASQ]-[PAS]-G-[PAD]-[FY]-x-[LI]-[DNQSTAP]-[DNH]-[LIVMFY].

NAME: SRP54-type proteins GTP-binding domain signature.

CONSENSUS: P-[LIVM]-x-[FYL]-[LIVMAT]-[GS]-x-[GS]-[EQ]-x(4)-[LIVMF].

50 NAME: Cytochrome c oxidase assembly factor COX10/ctaB/cyoE signature.

CONSENSUS: [ED]-x-D-x(2)-M-x-R-T-x(2)-R-x(4)-G.

NAME: Cyclin-dependent kinases regulatory subunits signature 1.

55 CONSENSUS: Y-S-x-[KR]-Y-x-[DE](2)-x-[FY]-E-Y-R-H-V-x-[LV]-[PT]-[KRP].

- NAME: Cyclin-dependent kinases regulatory subunits signature 2.
 CONSENSUS: H-x-P-E-x-H-[IV]-L-L-F-[KR].
- 5 NAME: Pentaxin family signature.
 CONSENSUS: H-x-C-x-[ST]-W-x-[ST].
- NAME: Immunoglobulins and major histocompatibility complex proteins signature.
 10 CONSENSUS: [FY]-x-C-x-[VA]-x-H.
- NAME: Prion protein signature 1.
 CONSENSUS: A-G-A-A-A-A-G-A-V-V-G-G-L-G-G-Y.
- 15 NAME: Prion protein signature 2.
 CONSENSUS: E-x-[ED]-x-K-[LIVM](2)-x-[KR]-[LIVM](2)-x-[QE]-M-C-x(2)-Q-Y.
- NAME: Cyclins signature.
 20 CONSENSUS: R-x(2)-[LIVMSA]-x(2)-[FYWS]-[LIVM]-x(8)-[LIVMFC]-x(4)-[LIVMFYA]-x(2)-
 CONSENSUS: [STAGC]-[LIVMFYQ]-x-[LIVMFYC]-[LIVMFY]-D-[RKH]-[LIVMFYW].
- 25 NAME: Proliferating cell nuclear antigen signature 1.
 CONSENSUS: [GA]-[LIVMF]-x-[LIVMA]-x-[SAV]-[LIVM]-D-x-[NSAE]-[HKR]-[VI]-x-[LY]-
 CONSENSUS: [VGAI]-x-[LIVM]-x-[LIVM]-x(4)-F.
- 30 NAME: Proliferating cell nuclear antigen signature 2.
 CONSENSUS: [RKA]-C-[DE]-[RH]-x(3)-[LIVMF]-x(3)-[LIVM]-x-[SGAN]-[LIVMF]-x-K-
 CONSENSUS: [LIVMF](2).
- 35 NAME: Actin-depolymerizing proteins signature.
 CONSENSUS: P-[DE]-x-[SA]-x-[LIVMT]-[KR]-x-[KR]-M-[LIVM]-[YA]-[STA](3)-x(3)-[LIVMF]-
 CONSENSUS: [KR].
- 40 NAME: BCL2-like apoptosis inhibitors (spans part of BH3, BH1 and BH2).
 NAME: Apoptosis regulator, Bcl-2 family BH1 domain signature.
 45 CONSENSUS: [LVME]-[FT]-x-[GSD]-[GL]-x(1,2)-[NS]-[YW]-G-R-[LIV]-[LIVC]-[GAT]-
 CONSENSUS: [LIVMF](2)-x-F-[GSAE]-[GSARY].
- NAME: Apoptosis regulator, Bcl-2 family BH2 domain signature.
 50 CONSENSUS: W-[LIM]-x(3)-[GR]-G-[WQ]-[DENSAV]-x-[FLGA]-[LIVFTC].
- NAME: Apoptosis regulator, Bcl-2 family BH3 domain signature.
 55 CONSENSUS: [LIVAT]-x(3)-L-[KARQ]-x-[IVAL]-G-D-[DESG]-[LIMFV]-[DENSHQ]-[LVSHRQ]-
 CONSENSUS: [NSR].

NAME: Apoptosis regulator, Bcl-2 family BH4 domain
 signature.
 5 CONSENSUS: [DS]-[NT]-R-[AE]-[LI]-V-x-[K]-[FY]-[LIV]-[GHS]-Y-
 K-L-[SR]-Q-[RK]-G-
 CONSENSUS: [HY]-x-[CW].

NAME: Apoptosis regulator, Bcl-2 family BH4 domain profile.

10 NAME: Arrestins signature.
 CONSENSUS: [FY]-R-Y-G-x-[DE](2)-x-[DE]-[LIVM](2)-G-[LIVM]-x-
 F-x-[RK]-[DEQ]-[LIVM].

15 NAME: AAA-protein family signature.
 CONSENSUS: [LIVMT]-x-[LIVMT]-[LIVMF]-x-[GATMC]-[ST]-[NS]-
 x(4)-[LIVM]-D-x-A-[LIFA]-
 CONSENSUS: x-R.

20 NAME: Ubiquitin domain signature.
 CONSENSUS: K-x(2)-[LIVM]-x-[DESAK]-x(3)-[LIVM]-[PA]-x(3)-Q-x-
 [LIVM]-[LIVMC]-
 CONSENSUS: [LIVMFY]-x-G-x(4)-[DE].

25 NAME: Ubiquitin domain profile.

NAME: ADP-ribosylation factors family signature.
 CONSENSUS: [HRQT]-x-[FYWI]-x-[LIVM]-x(4)-A-x(2)-G-x(2)-
 [LIVM]-x(2)-[GSA]-[LIVMF]-x-
 30 CONSENSUS: [WK]-[LIVM].

NAME: GTP-binding nuclear protein ran signature.
 CONSENSUS: D-T-A-G-Q-E-K-[LF]-G-G-L-R-[DE]-G-Y-Y.

35 NAME: SARL family signature.
 CONSENSUS: R-x-[LIVM]-E-V-F-M-C-S-[LIVM](2)-x-[KRQ]-x-G-Y-x-
 E-[AG]-[FI]-x-W-[LIVM]-
 CONSENSUS: x-Q-Y.

40 NAME: Band 7 protein family signature.
 CONSENSUS: R-x(2)-[LIV]-[SAN]-x(6)-[LIV]-D-x(2)-T-x(2)-W-G-
 [LIV]-[KRH]-[LIV]-x-
 CONSENSUS: [KR]-[LIV]-E-[LIV]-[KR].

45 NAME: Trp-Asp (WD) repeats signature.
 CONSENSUS: [LIVMSTAC]-[LIVMFYWSTAGC]-[LIMSTAG]-[LIVMSTAGC]-
 x(2)-[DN]-x(2)-
 CONSENSUS: [LIVMSTAC]-x-[LIVMFSTAG]-W-[DEN]-[LIVMFSTAGCN].

50 NAME: G-protein gamma subunit profile.

NAME: Ras GTPase-activating proteins signature.
 CONSENSUS: [GSN]-x-[LIVMF]-[FY]-[LIVMFY]-R-[LIVMFY](2)-
 [GACN]-P-[AV]-[LIV](2)-
 55 CONSENSUS: [SGAN]-P.

NAME: Ras GTPase-activating proteins profile.

NAME: Guanine-nucleotide dissociation stimulators CDC24
family signature.
CONSENSUS: L-x(2)-[LIVMFYW]-L-x(2)-P-[LIVM]-x(2)-[LIVM]-x-
[KRS]-x(2)-L-x-[LIVM]-x-
5 CONSENSUS: [DEQ]-[LIVM]-x(3)-[EST].

NAME: Guanine-nucleotide dissociation stimulators CDC25
family signature.
CONSENSUS: [GAP]-[CT]-V-P-[FY]-x(4)-[LIVMFY]-x-[DN]-[LIVM].

10 NAME: MARCKS family signature 1.
CONSENSUS: G-Q-E-N-G-H-V-[KR].

NAME: MARCKS family phosphorylation site domain.
15 CONSENSUS: E-T-P-K(5)-x(0,1)-F-S-F-K-K-x-F-K-L-S-G-x-S-F-K-
[KR]-[NS]-[KR]-K-E.

NAME: Stathmin family signature 1.
CONSENSUS: P-[KQ]-[KR](2)-[DE]-x-S-L-[EG]-E.

20 NAME: Stathmin family signature 2.
CONSENSUS: A-E-K-R-E-H-E-[KR]-E-V.

NAME: GTP-binding elongation factors signature.
25 CONSENSUS: D-[KRSTGANQFYW]-x(3)-E-[KRAQ]-x-[RKQD]-[GC]-
[IVMK]-[EST]-[IV]-x(2)-
CONSENSUS: [GSTACKRNQ].

NAME: Elongation factor 1 beta/beta'/delta chain signature
30 1.
CONSENSUS: [DE]-[DEG]-[DE](2)-[LIVMF]-D-L-F-G.

NAME: Elongation factor 1 beta/beta'/delta chain signature
2.
35 CONSENSUS: V-Q-S-x-D-[LIVM]-x-A-[FWM]-[NQ]-K-[LIVM].

NAME: Elongation factor 1 gamma chain profile.

NAME: Elongation factor Ts signature 1.
40 CONSENSUS: L-R-x(2)-T-[GDQ]-x-[GS]-[LIVMF]-x(0,1)-[DENKAC]-x-
K-[KRNEQS]-[AV]-L.

NAME: Elongation factor Ts signature 2.
CONSENSUS: E-[LIVM]-N-[SCV]-[QE]-T-D-F-V-[SA]-[KRN].

NAME: Elongation factor P signature.
CONSENSUS: K-x-A-x(4)-G-x(2)-[LIV]-x-V-P-x(2)-[LIV]-x(2)-G.

NAME: Eukaryotic initiation factor 1A signature.
50 CONSENSUS: [IM]-x-G-x-[GS]-[KRH]-x(4)-[CL]-x-D-G-x(2)-R-x(2)-
[RH]-I-x-G.

NAME: Eukaryotic initiation factor 4E signature.
CONSENSUS: [DE]-[IFY]-x(2)-F-[KR]-x(2)-[LIVM]-x-P-x-W-E-[DV]-
55 x(5)-G-G-[KR]-W.

NAME: Eukaryotic initiation factor 5A hypusine signature.
CONSENSUS: [PT]-G-K-H-G-x-A-K.

- NAME: Initiation factor 2 signature.
 CONSENSUS: G-x-[LIVM]-x(2)-L-[KR]-[KRHNS]-x-K-x(5)-[LIVM]-
 x(2)-G-x-[DEN]-C-G.
- 5 NAME: Initiation factor 3 signature.
 CONSENSUS: [KR]-[LIVM](2)-[DN]-[FY]-[GSN]-[KR]-[LIVMFYS]-x-
 [FY]-[DET]-x(2)-[KR].
- 10 NAME: Translation initiation factor SUI1 signature.
 CONSENSUS: [LIVM]-[EQ]-[LIVM]-Q-G-[DEN]-[KHQ]-[KRV].
- NAME: Prokaryotic-type class I peptide chain release factors
 signature.
 15 CONSENSUS: [AR]-[STA]-x-G-x-G-G-Q-[HNGCS]-V-N-x(3)-[ST]-A-
 [IV].
- NAME: Transcription termination factor nusG signature.
 CONSENSUS: [LIVM]-F-G-[KRW]-x-T-P-[IV]-x-[LIVM].
- 20 NAME: Calponin family repeat.
 CONSENSUS: [LIVM]-x-[LS]-Q-[MAS]-G-[STY]-[ENT]-[KRQ]-x(2)-
 [STN]-Q-x-G-x(3,4)-G.
- 25 NAME: CAP protein signature 1.
 CONSENSUS: [LIVM](2)-x-R-L-[DE]-x(4)-R-L-E.
- NAME: CAP protein signature 2.
 CONSENSUS: D-[LIVMFY]-x-E-x-[PA]-x-P-E-Q-[LIVMFY]-K.
- 30 NAME: Calreticulin family signature 1.
 CONSENSUS: [KRHN]-x-[DEQN]-[DEQNK]-x(3)-C-G-G-[AG]-[FY]-
 [LIVM]-[KN]-[LIVMFY](2).
- 35 NAME: Calreticulin family signature 2.
 CONSENSUS: [LIVM](2)-F-G-P-D-x-C-[AG].
- NAME: Calreticulin family repeated motif signature.
 CONSENSUS: [IV]-x-D-x-[DENST]-x(2)-K-P-[DEH]-D-W-[DEN].
- 40 NAME: Calsequestrin signature 1.
 CONSENSUS: [EQ]-[DE]-G-L-[DN]-F-P-x-Y-D-G-x-D-R-V.
- NAME: Calsequestrin signature 2.
 45 CONSENSUS: [DE]-L-E-D-W-[LIVM]-E-D-V-L-x-G-x-[LIVM]-N-T-E-D-
 D-D.
- NAME: S-100/ICaBP type calcium binding protein signature.
 CONSENSUS: [LIVMFYW](2)-x(2)-[LK]-D-x(3)-[DN]-x(3)-[DNSG]-
 [FY]-x-[ES]-[FYVC]-x(2)-
- 50 CONSENSUS: [LIVMFS]-[LIVMF].
- NAME: Hemolysin-type calcium-binding region signature.
 CONSENSUS: D-x-[LI]-x(4)-G-x-D-x-[LI]-x-G-G-x(3)-D.
- 55 NAME: HlyD family secretion proteins signature.
 CONSENSUS: [LIVM]-x(2)-G-[LM]-x(3)-[STGAV]-x-[LIVMT]-x-
 [LIVMT]-[GE]-x-[KR]-x-

CONSENSUS: [LIVMFYW](2)-x-[LIVMFYW](3).

NAME: P-II protein urydylatlon site.

CONSENSUS: Y-[KR]-G-[AS]-[AE]-Y.

5

NAME: P-II protein C-terminal region signature.

CONSENSUS: [ST]-x(3)-G-[DY]-G-[KR]-[IV]-[FW]-[LIVM]-x(2)-[LIVM].

10 NAME: 14-3-3 proteins signature 1.

CONSENSUS: R-N-L-[LIV]-S-[VG]-[GA]-Y-[KN]-N-[IVA].

NAME: 14-3-3 proteins signature 2.

15 CONSENSUS: Y-K-[DE]-S-T-L-I-[IM]-Q-L-[LF]-[RHC]-D-N-[LF]-T-[LS]-W-[TAN]-[SAD].

NAME: ATP1G1 / PLM / MAT8 family signature.

CONSENSUS: [DNS]-x-F-x-Y-D-x(2)-[ST]-[LIVM]-[RQ]-x(2)-G.

20 NAME: BTG1 family signature 1.

CONSENSUS: Y-x(2)-[HP]-W-[FY]-[AP]-E-x-P-x-K-G-x-[GA]-[FY]-R-C-[IV]-[RH]-[IV].

NAME: BTG1 family signature 2.

25 CONSENSUS: [LV]-P-x-[DE]-[LM]-[ST]-[LIVM]-W-[IV]-D-P-x-E-V-[SC]-x-[RQ]-x-G-E.

NAME: Cullin family signature.

30 CONSENSUS: [LIV]-K-x(2)-[LIV]-x(2)-L-I-[DEQ]-[KRHNQ]-x-Y-[LIVM]-x-R-x(6,7)-[FY]-x-
CONSENSUS: Y-x-[SA]>.

NAME: Cullin family profile.

35 NAME: Enhancer of rudimentary signature.

CONSENSUS: Y-D-I-[SA]-x-L-[FY]-x-F-[IV]-D-x(3)-D-[LIV]-S.

NAME: G10 protein signature 1.

40 CONSENSUS: L-C-C-x-[KR]-C-x(4)-[DE]-x-N-x(4)-C-x-C-R-V-P.

NAME: G10 protein signature 2.

CONSENSUS: C-x-H-C-G-C-[KRH]-G-C-[SA].

NAME: Glucokinase regulatory protein family signature.

45 CONSENSUS: G-[PA]-E-x-[LIV]-[STA]-G-S-[ST]-R-[LIVM]-K-[STGA](3)-x(2)-K.

NAME: GTP1/0BG family signature.

50 CONSENSUS: D-[LIVM]-P-G-[LIVM](2)-[DEY]-[GN]-A-x(2)-G-x-G.

NAME: HIT family signature.

CONSENSUS: [NQA]-x(4)-[GAV]-x-[QF]-x-[LIVM]-x-H-[LIVMFYT]-H-[LIVMFT]-H-[LIVMF](2)-
CONSENSUS: [PSGA].

55

NAME: Caseins alpha/beta signature.

CONSENSUS: C-L-[LV]-A-x-A-[LVF]-A.

- NAME: Clathrin adaptor complexes medium chain signature 1.
 CONSENSUS: [IVT]-[GSP]-W-R-x(2,3)-[GAD]-x(2)-[HY]-x(2)-N-x-
 [LIVMAFY](3)-D-[LIVM]-
 CONSENSUS: [LIVMT]-E.
- 5 NAME: Clathrin adaptor complexes medium chain signature 2.
 CONSENSUS: [LIV]-x-F-I-P-P-x-G-x-[LIVMFY]-x-L-x(2)-Y.
- 10 NAME: Clathrin adaptor complexes small chain signature.
 CONSENSUS: [LIVM](2)-Y-[KR]-x(4)-L-Y-F.
- NAME: Ependymins signature 1.
 CONSENSUS: F-E-E-G-x-[LIVMF]-Y-[ED]-I-D-x(2)-N-[QE]-S-C-
 [RKH](2).
- 15 NAME: Ependymins signature 2.
 CONSENSUS: [QE]-[LIVMA]-F-x(2)-P-[STA]-[FY]-C-[DE]-[GA]-
 [LIVM]-x(2)-[DE](2).
- 20 NAME: Syntaxin / epimorphin family signature.
 CONSENSUS: [RQ]-x(3)-[LIVMA]-x(2)-[LIVM]-[ESH]-x(2)-[LIVMT]-
 x-[DEV]-[LIVM]-x(2)-
 CONSENSUS: [LIVM]-[FS]-x(2)-[LIVM]-x(3)-[LIVT]-x(2)-Q-
 [GADEQ]-x(2)-[LIVM]-[DNQT]-x-
 25 CONSENSUS: [LIVMF]-[DES]-x(2)-[LIVM].
- NAME: Extracellular proteins SCP/Tpx-1/Ag5/PR-1/Sc7
 signature 1.
 CONSENSUS: [GDER]-H-[FYWH]-T-Q-[LIVM](2)-W-x(2)-[STN].
- 30 NAME: Extracellular proteins SCP/Tpx-1/Ag5/PR-1/Sc7
 signature 2.
 CONSENSUS: [LIVMFYH]-[LIVMFY]-x-C-[NQRHS]-Y-x-[PARH]-x-[GL]-
 N-[LIVMFYWDN].
- 35 NAME: Fetuin family signature 1.
 CONSENSUS: C-x(56)-C-x(10)-C-x(13)-C-x(17,18)-C-x(13)-C-x(2)-
 C-x(58)-C-x(10,11)-
 CONSENSUS: C-x(10,12)-C-x(16,22)-C.
- 40 NAME: Fetuin family signature 2.
 CONSENSUS: L-E-T-x-C-H-x-L-D-P-T-P.
- NAME: Legume lectins beta-chain signature.
 45 CONSENSUS: [LIV]-[STAG]-V-[DEQV]-[FLI]-D-[ST].
- NAME: Legume lectins alpha-chain signature.
 CONSENSUS: [LIV]-x-[EDQ]-[FYWKR]-V-x-[LIV]-G-[LF]-[ST].
- 50 NAME: Vertebrate galactoside-binding lectin signature.
 CONSENSUS: W-[GEK]-x-[EQ]-x-[KRE]-x(3,6)-[PCTF]-[LIVMF]-
 [NQEKGSKV]-x-[GH]-x(3)-
 CONSENSUS: [DENKHS]-[LIVMFC].
- 55 NAME: Lysosome-associated membrane glycoproteins duplicated
 domain signature.
 CONSENSUS: [STA]-C-[LIVM]-[LIVMFYW]-A-x-[LIVMFYW]-x(3)-
 [LIVMFYW]-x(3)-Y.

- NAME: LAMP glycoproteins transmembrane and cytoplasmic domain signature.
 5 CONSENSUS: C-x(2)-D-x(3,4)-[LIVM](2)-P-[LIVM]-x-[LIVM]-G-x(2)-[LIVM]-x-G-[LIVM](2)-
 CONSENSUS: x-[LIVM](4)-A-[FY]-x-[LIVM]-x(2)-[KR]-[RH]-x(1,2)-[STAG](2)-Y-[EQ].
- NAME: Glycophorin A signature.
 10 CONSENSUS: I-I-x-[GAC]-V-M-A-G-[LIVM](2).
- NAME: PMP-22 / EMP / MP20 family signature 1.
 CONSENSUS: [LIVMF](4)-[SA]-T-x(2)-[DNKS]-x-W-x(9,13)-[LIV]-W-x(2)-C.
 15
- NAME: PMP-22 / EMP / MP20 family signature 2.
 CONSENSUS: [RQ]-[AV]-x-M-[IV]-L-S-x-[LI]-x(4)-[GSA]-[LIVMF](3).
 20
- NAME: Oxysterol-binding protein family signature.
 CONSENSUS: E-[KQ]-x-S-H-[HR]-P-P-x-[STACF]-A.
- NAME: Yeast PIR proteins repeats signature.
 25 CONSENSUS: S-Q-[IV]-[STGNH]-D-G-Q-[LIV]-Q-[AIV]-[STA].
- NAME: Seminal vesicle protein I repeats signature.
 CONSENSUS: [IVM]-x-G-Q-D-x-V-K-x(5)-[KN]-G-x(3)-[STLV].
 30
- NAME: Seminal vesicle protein II repeats signature.
 CONSENSUS: [GSA]-Q-x-K-S-[FY]-x-Q-x-K-[SA].
- NAME: Serum amyloid A proteins signature.
 CONSENSUS: A-R-G-N-Y-[ED]-A-x-[QKR]-R-G-x-G-G-x-W-A.
- NAME: Spermadhesins family signature 1.
 35 CONSENSUS: C-G-x(2)-[LI]-x(4)-G-x-I-x(9)-C-x-W-T.
- NAME: Spermadhesins family signature 2.
 40 CONSENSUS: C-x-K-E-x-[LIVM]-E-[LIVM]-x-[DE]-x(3)-[GS]-x(5)-K-x-C.
- NAME: Stress-induced proteins SRP1/TIP1 family signature.
 CONSENSUS: P-W-Y-[ST](2)-R-L.
- NAME: Glypicans signature.
 45 CONSENSUS: C-x(2)-C-x-G-[LIVM]-x(4)-P-C-x(2)-[FY]-C-x(2)-[LIVM]-x(2)-G-C.
- NAME: Syndecans signature.
 50 CONSENSUS: [FY]-R-[IM]-[KR]-K(2)-D-E-G-S-Y.
- NAME: Tissue factor signature.
 CONSENSUS: W-K-x-K-C-x(2)-T-x-[DEN]-T-E-C-D-[LIVM]-T-D-E.
- NAME: Translationally controlled tumor protein signature 1.
 55 CONSENSUS: [IA]-G-[GAS]-N-[PA]-S-A-E-[GDE]-[PAGE]-x(0,1)-[DEG]-x-[DEN]-x(2)-[DE].

- NAME: Translationally controlled tumor protein signature 2.
 CONSENSUS: [FL]-[FY]-[IVT]-G-E-x-[MA]-x(2,5)-[DEN]-[GAS]-x-
 [LV]-[AV]-x(3)-[FY]-[KR]-
 CONSENSUS: [DE].
- 5 NAME: Tub family signature 1.
 CONSENSUS: F-[KHQ]-G-R-V-[ET]-x-A-S-V-K-N-F-Q.
- 10 NAME: Tub family signature 2.
 CONSENSUS: A-F-[AG]-I-[SAC]-[LIVM]-[ET]-S-F-x-[GST]-K-x-A-C-E.
- NAME: HCP repeats signature.
 CONSENSUS: H-R-H-R-G-H-x(2)-[DE](7).
- 15 NAME: Bacterial ice-nucleation proteins octamer repeat.
 CONSENSUS: A-G-Y-G-S-T-x-T.
- NAME: Cell cycle proteins ftsW / rodA / spoVE signature.
 20 CONSENSUS: [NV]-x(5)-[GTR]-[LIVMA]-x-P-[PTLIVM]-x-G-[LIVM]-
 x(3)-[LIVFW](2)-S-[YSA]-
 CONSENSUS: G-G-[ETN]-[SA].
- NAME: Enterobacterial virulence outer membrane protein
 25 signature 1.
 CONSENSUS: G-[LIVMFY]-N-[LIVM]-K-Y-R-Y-E.
- NAME: Enterobacterial virulence outer membrane protein
 30 signature 2.
 CONSENSUS: [FYW]-x(2)-G-x-G-Y-[KR]-F>.
- NAME: Hydrogenases expression/synthesis hypA family
 signature.
 CONSENSUS: F-[CSA]-[FY]-[DE]-[LIVA](2)-x(3)-[ET]-[LIVM]-
 35 x(1b)-C-x(2)-C-x(12,15)-
 CONSENSUS: C-P-x-C.
- NAME: Hydrogenases expression/synthesis hupF/hypC family
 signature.
 40 CONSENSUS: <M-C-[LIV]-[GA]-[LIV]-P-x-[QKR]-[LIV].
- NAME: Staphylocoagulase repeat signature.
 CONSENSUS: A-R-P-x(3)-K-x-S-x-T-N-A-Y-N-V-T-T-x(2)-[DN]-G-
 45 x(3)-Y-G.
- NAME: 11-S plant seed storage proteins signature.
 CONSENSUS: N-G-x-[DE](2)-x-[LIVMF]-C-[ET]-x(11,12)-[PAG]-D.
- NAME: Dehydrins signature 1.
 50 CONSENSUS: S(5)-[DE]-x-[DE]-G-x(1,2)-G-x(0,1)-[KR](4).
- NAME: Dehydrins signature 2.
 CONSENSUS: [KR]-[LIM]-K-[DE]-K-[LIM]-P-G.
- 55 NAME: Germin family signature.
 CONSENSUS: G-x(4)-H-x-H-P-x-A-x-E-[LIVM].
- NAME: Oleosins signature.

CONSENSUS: [EAG]-[EST]-x(2)-[EAG]-x(2)-[LIVM]-[ESAD]-T-P-
 [LIVMF](4)-F-S-P-[LIVM](3)-
 CONSENSUS: P-A.

5 NAME: Small hydrophilic plant seed proteins signature.
 CONSENSUS: G-[EQ]-T-V-V-P-G-G-T.

NAME: Pathogenesis-related proteins BetvI family signature.
 CONSENSUS: G-x(2)-[LIVMF]-x(4)-E-x(2)-[CSTAEN]-x(8,9)-[GND]-
 10 G-[GS]-[CS]-x(2)-K-x(4)-
 CONSENSUS: [FY].

NAME: Pollen proteins Ole e I family signature.
 CONSENSUS: [EQ]-G-x-V-Y-C-D-T-C-R.

15 NAME: Thaumatin family signature.
 CONSENSUS: G-x-[GF]-x-C-x-T-[GA]-D-C-x(1,2)-G-x(2,3)-C.

NAME: Mrp family signature.
 20 CONSENSUS: W-x(2)-[LIVM]-D-[LIVMY](4)-D-x-P-P-G-T-[GS]-D.

NAME: Glucose inhibited division protein A family signature
 1.
 CONSENSUS: [GS]-P-x-Y-C-P-S-[LIVM]-E-x-K-[LIVM]-x-[KR]-F.

25 NAME: Glucose inhibited division protein A family signature
 2.
 CONSENSUS: A-G-Q-x-[ENT]-G-x(2)-G-Y-x-E-[SAG](3)-[QS]-G-
 [LIVM](2)-A-G-[LIVMT]-N-A.

30 NAME: NOL1/NOP2/sun family signature.
 CONSENSUS: [FV]-D-[KRA]-[LIVMA]-L-x-D-[AV]-P-C-[EST]-[GA].

NAME: PET112 family signature.
 35 CONSENSUS: [DN]-x-[DN]-R-x(3)-P-L-[LIV]-E-[LIV]-x-[EST]-x-P.

NAME: Protein smpB signature.
 CONSENSUS: [TA]-G-[LIVM]-x-L-x-G-x-E-[LIVM]-[KQ]-[SA]-[LIVM].

40 NAME: Hypothetical cof family signature 1.
 CONSENSUS: [LIVFYAN]-[LIVMF]-x(2)-D-[LIVMF]-[ND]-G-T-[LIV]-
 [LVY]-[STANLM].

NAME: Hypothetical cof family signature 2.
 45 CONSENSUS: [LIVMFC]-G-D-[GSANQ]-x-N-D-x(3)-[LIMFY]-x(2)-[AV]-
 x(2)-[GSCP]-x(2)-
 CONSENSUS: [LMP]-x(2)-[GAS].

NAME: RI01/ZK632.3/MJ0444 family signature.
 50 CONSENSUS: [LIVM]-V-H-[GA]-D-L-S-E-[FY]-N-x-[LIVM].

NAME: SUA5/yci0/yrdC family signature.
 CONSENSUS: [LIVMTA](3)-[LIVMFYC]-[PG]-T-[DE]-[STA]-x-[FY]-
 [GA]-[LIVM]-[GS].

55 NAME: Uncharacterized protein family UPF0001 signature.
 CONSENSUS: [FW]-H-[FM]-[IV]-G-x-[LIV]-Q-x-[NKR]-K-x(3)-[LIV].

- NAME: Uncharacterized protein family UPF0003 signature.
 CONSENSUS: G-x-V-x(2)-[LIV]-x(3)-[SA]-x(6)-D-x(3)-[LIVT](3)-
 P-N-x(2)-[LIVMF](2)-
 CONSENSUS: x(5)-N.
- 5 NAME: Uncharacterized protein family UPF0004 signature.
 CONSENSUS: [LIVM]-x-[LIVMT]-x(2)-G-C-x(3)-C-[STAN]-[FY]-C-x-
 [LIVM]-x(4)-G.
- 10 NAME: Uncharacterized protein family UPF0005 signature.
 CONSENSUS: G-[LIVM](2)-[SA]-x(5,8)-G-x(2)-[LIVM]-G-P-x-L-
 x(4)-[SAG]-x(4,6)-
 CONSENSUS: [LIVM](2)-x(2)-A-x(3)-T-A-[LIVM](2)-F.
- 15 NAME: Uncharacterized protein family UPF0006 signature 1.
 CONSENSUS: [LIVMFY](2)-D-[STA]-H-x-H-[LIVMF]-[DN].
- NAME: Uncharacterized protein family UPF0006 signature 2.
 CONSENSUS: P-[LIVM]-x-[LIVM]-H-x-R-x-[TA]-x-[DE].
- 20 NAME: Uncharacterized protein family UPF0006 signature 3.
 CONSENSUS: [LVSA]-[LIVA]-x(2)-[LIVM]-[PS]-x(3)-L-[LIVM]-
 [LIVMS]-E-T-D-x-P.
- 25 NAME: Uncharacterized protein family UPF0007 signature.
 CONSENSUS: V-L-[IV]-H-D-[GA]-A-R.
- NAME: Uncharacterized protein family UPF0011 signature.
 CONSENSUS: S-D-A-G-x-P-x-[LIV]-[SN]-D-P-G.
- 30 NAME: Uncharacterized protein family UPF0012 signature.
 CONSENSUS: [GTA]-x(2)-[IVT]-C-Y-D-[LIVM]-x-F-P-x(9)-G.
- NAME: Uncharacterized protein family UPF0015 signature.
 CONSENSUS: [DE]-[LIVMF](3)-R-T-[SG]-G-x(2)-R-x-S-x-[FY]-
 [LIVM](2)-W-Q.
- 35 NAME: Uncharacterized protein family UPF0016 signature.
 CONSENSUS: E-[LIVM]-G-D-K-T-F-[LIVMF](2)-A.
- 40 NAME: Uncharacterized protein family UPF0017 signature.
 CONSENSUS: D-x(8)-[GN]-[LFY]-x(4)-[DET]-[LY]-Y-x(3)-[ST]-
 x(7)-[IV]-x(2)-[PS]-x-
 CONSENSUS: [LIVM]-x-[LIVM]-x(3)-[DN]-D.
- 45 NAME: Uncharacterized protein family UPF0019 signature.
 CONSENSUS: L-P-V-[VT]-[NQL]-F-[AT]-A-G-G-[LIV]-A-T-P-A-D-A-A-
 [LM].
- 50 NAME: Uncharacterized protein family UPF0020 signature.
 CONSENSUS: D-P-[LIVMF]-C-G-[ST]-G-x(3)-[LI]-E.
- NAME: Uncharacterized protein family UPF0021 signature.
 CONSENSUS: C-K-x(2)-F-x(4)-E-x(22,23)-S-G-G-K-D.
- 55 NAME: Uncharacterized protein family UPF0023 signature.
 CONSENSUS: D-x-D-E-[LIV]-L-x(4)-V-F-x(3)-S-K-G.

- NAME: Uncharacterized protein family UPF0024 signature.
 CONSENSUS: G-x-K-D-[K R]-x-A-[L V]-T-x-Q-x-[L I V F]-[S G C].
- 5 NAME: Uncharacterized protein family UPF0025 signature.
 CONSENSUS: D-V-[L I V]-x(2)-G-H-[E S T]-H-x(12)-[L I V M F]-N-P-G.
- NAME: Uncharacterized protein family UPF0027 signature.
 CONSENSUS: Q-[L I V M]-x-N-x-A-x-[L I V M]-P-x-I-x(6)-[L I V M]-P-D-x-H-x-G-x-G-x(2)-[L I V]-G.
- 10 NAME: Uncharacterized protein family UPF0028 signature.
 CONSENSUS: [E G A]-[E G S]-G-[E G A]-A-R-G-x-[S A]-H-x-G-x(9)-[L I V]-x-[L I V]-D-x(2)-[E G A]-G-x-S-
 CONSENSUS: x-G.
- 15 NAME: Uncharacterized protein family UPF0029 signature.
 CONSENSUS: G-x(2)-[L I V M](2)-x(2)-[L I V M]-x(4)-[L I V M]-x(5)-[L I V M](2)-x-R-[F Y W](2)-G-
 CONSENSUS: G-x(2)-[L I V M]-G.
- 20 NAME: Uncharacterized protein family UPF0030 signature.
 CONSENSUS: [E G A]-L-I-[L I V]-P-G-G-E-S-T-[E S T A].
- NAME: Uncharacterized protein family UPF0031 signature 1.
 25 CONSENSUS: [E S A V]-[L I V W]-[L V A]-[L I V]-G-[P N S]-G-L-[E G P]-x-[E D E N Q T].
- NAME: Uncharacterized protein family UPF0031 signature 2.
 CONSENSUS: [E G A]-G-x-G-D-[T V]-[E L T]-[E S T A]-G-x-[L I V M].
- 30 NAME: Uncharacterized protein family UPF0032 signature.
 CONSENSUS: Y-x(2)-F-[L I V M A](2)-x-L-x(4)-G-x(2)-F-[E Q]-[L I V M F]-P-[L I V M].
- NAME: Uncharacterized protein family UPF0033 signature.
 35 CONSENSUS: L-[D N]-x(2)-[T A G]-x(2)-C-P-x-P-x-[L I V M].
- NAME: Uncharacterized protein family UPF0034 signature.
 CONSENSUS: [L I V M]-[D N G]-[L I V M]-N-x-G-C-P-x(3)-[L I V M A S Q]-x(5)-G-[E S A C].
- 40 NAME: Uncharacterized protein family UPF0035 signature.
 CONSENSUS: L-L-T-x-R-[S A]-x(3)-R-x(3)-G-x(3)-F-P-G-G.
- NAME: Uncharacterized protein family UPF0036 signature.
 45 CONSENSUS: H-x-S-G-H-[E G A]-x(3)-[E D E]-x(3)-[L M]-x(5)-P-x(3)-[L I V M]-P-x-H-G-[E D E].
- NAME: Uncharacterized protein family UPF0038 signature.
 50 CONSENSUS: G-x-[L I]-x-R-x(2)-L-x(4)-F-x(8)-[L I V]-x(5)-P-x-[L I V].
- NAME: Uncharacterized protein family UPF0044 signature.
 CONSENSUS: L-[E S T]-x(3)-K-x(3)-[K R]-[E S G A]-x-[E G A]-H-x-L-x-P-[L I V]-x(2)-[L I V]-[E G A]-
 55 CONSENSUS: x(2)-G.
- NAME: Uncharacterized protein family UPF0047 signature.

CONSENSUS: S-X(2)-[LIV]-x-[LIV]-x(2)-G-x(4)-G-T-W-Q-x-[LIV].

NAME: Uncharacterized protein family UPF0054 signature.

CONSENSUS: H-[GS]-x-L-H-L-[LI]-G-[FYW]-D-H.

5

NAME: Uncharacterized protein family UPF0057 signature.

CONSENSUS: [LIV]-x-[STA]-[LIVF](3)-P-P-[LIVA]-[GA]-[IV]-x(4)-[GKN].

10

NAME: Hypothetical YER057c/yjjV family signature.

CONSENSUS: P-[AT]-R-[SA]-x-[LIVMY]-x(2)-[AK]-x-L-P-x(4)-[LIVM]-E.

NAME: Hypothetical hesB/yadR/yfhF family signature.

15

CONSENSUS: F-x-[LIVMFY]-x-N-[PG]-[NSK]-x(4)-C-x-C-[GS]-x-S-F.

NAME: Hypothetical yab0/yceC/sfhB family signature.

CONSENSUS: [NHY]-R-[LI]-D-x(2)-T-[ST]-G-[LIVMA]-[LIVMF](2)-[LIVMFG]-[SGAC].

20

Deposit of Clones

25

Each clone has been transfected into separate bacterial cells (E. coli) in the composite deposit.

The clones are located and publically available from the Resource Center of the German Human Genome Project (Heubner Weg 6, 14059 Berlin, GERMANY), from which each clone comprising a particular polynucleotide is obtainable. The Resource Center library numbers are slightly different than those presented here, but may be readily obtained by the following key or with the assistance of Resource Center personnel.

30

The library name becomes a number: brain (hfbr2) becomes 5b4; kidney (hfkd2) becomes 5bb; mammary carcinoma (hmcfl) becomes 727; testis (htes3) becomes 434; amygdala (hamy2) becomes 7b1; melanoma (hmel2) becomes 7b2 and uterus (hutel) becomes 58b.

Next, the plate number is converted to two digits (e.g., "2" becomes "02") and is moved behind the plate coordinate, and the

40

underscore is dropped. The following examples are helpful:

	<u>Listed Number</u>	<u>Resource Center Number</u>
	DKFZphamy2_10h17	DKFZp7b1H1710
	DKFZphfbr2_78i21	DKFZp5b4I2178
45	DKFZphfkd2_3k1	DKFZp5bbK013
	DKFZphmcf1_1c23	DKFZp727C231
	DKFZhmel2_12j1	DKFZp7b2J0112
	DKFZphtes3_1bb5	DKFZp434B051b
	DKFZphutel_17k7	DKFZp58bK0717

The libraries were constructed using two commercially available vectors. The brain (hfbr2 designations) and kidney (hfk2 designations) libraries utilize pAMP 1 from Life Technologies and are maintained in XL-2Blue (Stratagene); the amygdala (hamy2), testes (htes3) and melanoma (hmel2) libraries are constructed in pSPORT1, also from Life Technologies, and are maintained in DH10B (Life Technologies). In addition to the following techniques, consultation with the commercial literature available on these clones will make evident all of the housekeeping techniques needed to propagate and isolate the individual constructs. All inserts may be excised with a NotI/SalI digestion. Alternatively, universal primers, flanking the cloning region, may be used to amplify the inserts using PCR methods.

Bacterial cells containing a particular clone can be obtained from the composite deposit as follows:

An oligonucleotide probe or probes should be designed to the sequence that is known for that particular clone. This sequence can be derived from the sequences provided herein, or from a combination of those sequences. Methods of probe design are presented below.

Oligonucleotide probes may be labeled with $-^{32}\text{P}$ ATP (specific activity 6000 Ci/mmol) and T4 polynucleotide kinase using commonly employed techniques for labeling oligonucleotides. Other, non-radioactive labeling techniques can also be used. Unincorporated label typically is removed by gel filtration chromatography or other established methods. The amount of radioactivity incorporated into the probe can be quantified by measurement in a scintillation counter. Preferably, specific activity of the resulting probe generally should be approximately 4×10^6 dpm/pmol.

The bacterial culture containing the pool of full-length clones should preferably be thawed and 100 μ l of the stock used to inoculate a sterile culture flask containing 25 ml of sterile L-broth containing ampicillin at 50 - 100 g/ml (for XL-2Blue strains 25 g/ml tetracycline should also be used). The culture should preferably be grown to saturation at 37°C., and the saturated culture should preferably be diluted in fresh L-broth.

Aliquots of these dilutions should preferably be plated to determine the dilution and volume which will yield approximately 5000 distinct and well-separated colonies on solid bacteriological media containing L-broth containing ampicillin at 100 g/ml (for XL-2Blue strains 25 g/ml tetracycline should also be used) and agar at 1.5% in a 150 mm petri dish when grown overnight at 37°C. Other known methods of obtaining distinct, well-separated colonies can also be employed.

Standard colony hybridization procedures should then be used to transfer the colonies to nitrocellulose filters and lyse, denature and bake them. The filter is then preferably incubated at 65°C. for 1 hour with gentle agitation in 6 x SSC (20 x stock is 175.3 g NaCl/liter, 88.2 g Na citrate/liter, adjusted to pH 7.0 with NaOH) containing 0.5% SDS, 100 g/ml of yeast RNA, and 10 mM EDTA (approximately 10 mL per 150 mm filter). Preferably, the probe is then added to the hybridization mix at a concentration greater than or equal to 1×10^6 dpm/mL. The filter is then preferably incubated at 65°C. with gentle agitation overnight. The filter is then preferably washed in 500 mL of 2 x SSC/0.5% SDS at room temperature without agitation, preferably followed by 500 mL of 2 x SSC/0.1% SDS at room temperature with gentle shaking for 15 minutes. A third wash with 0.1 x SSC/0.5% SDS at 65°C. for 30 minutes to 1 hour is optional. The filter is then preferably dried and subjected to autoradiography for sufficient time to visualize the positives on the X-ray film. Other known hybridization methods can also be employed.

The positive colonies are picked, grown in culture, and plasmid DNA isolated using standard procedures. The clones can then be verified by restriction analysis, hybridization analysis, or DNA sequencing.

Alternatively, clones may be grown as described above, and PCR used to isolate the insert DNAs. Methods of PCR are described below and are otherwise well known.

ERROR SCREENING

The DNA sequences found herein derive from individual clones, which are publicly available, as noted above. Thus, the skilled artisan will recognize that any specific sequence disclosed herein

readily can be screened for errors by resequencing a particular fragment, in both directions (i.e., by sequencing both strands). Alternatively, error screening can be performed by amplifying and/or cloning any of the inventive DNAs, using for example RT-PCR, and sequencing the resulting amplified product. In the event that there is a sequencing error, reference should be made to the deposited clone as the correct sequence.

USES AND BIOLOGICAL ACTIVITIES OF THE INVENTIVE MOLECULES

The inventive molecules and their derivatives are susceptible to a wide variety of uses, based on functional and/or structural properties. The skilled worker will appreciate, based on the biological activities detailed below, and discussed with regard to the individual sequences herein, that the inventive molecules will find usefulness in numerous therapeutic and diagnostic applications.

The DNA molecules, especially the potassium salts thereof, can be used as fertilizer supplements due to their high nitrogen and phosphorus contents. Since the DNAs are of defined length, they are also useful in gel electrophoresis as molecular weight markers. Due to their similarity with known molecules, certain of the DNA molecules and their variants and derivatives may be used in any number of different diagnostic procedures and therapeutic applications. They may also be used to make the encoded proteins.

The proteins themselves have many possible uses. They may be used as a nutritional supplement for humans, animals and even for laboratory use as, for example, medium for bacterial cultures. Moreover, since the proteins are of defined, known sizes, they may be used as molecular weight markers for gel electrophoresis and gel filtration. Because they are of defined sequences, they also have use in microsequencing and protein fingerprinting applications.

Expression Profiling Applications

Given their known tissue expression and functional associations, assemblages of the inventive proteins (or corresponding antibodies) and nucleic acids are particularly suited to expression profiling applications. Expression profiling generally entails constructing an array of indicators that signal

the presence of a particular RNA or protein expression product. Such arrays can be used to evaluate, for example, pharmacological effectiveness and toxicity. In particular, expression profiles from such arrays can be generated from cells treated with known compounds, having known properties, and these profiles can be compared to profiles of unknowns to evaluate similarities and differences, which can be correlated with efficacy or toxicity.

Additional uses of profiling include diagnosis, tracking development, and ascertaining signaling and metabolic pathways.

For examples of references describing profiling and its uses, see Farr et al., U.S. Patent 5,811,231 (1998); Seilhamer et al., U.S. Patent 5,840,484 (1998); Rine et al., U.S. Patent No. 5,777,888 (1998); WO 97/27317; WO 99/05323; WO 99/09218; and WO 99/14369. For a device for implementing such techniques, see Lipshutz et al., U.S. Patent No. 5,856,174 (1999) and Anderson et al., U.S. Patent No. 5,922,591 (1999).

In one embodiment, a subset of the inventive DNAs will be arrayed on a substrate, like a gene chip, a filter or a 96-well plate. Test samples containing cells are maintained in the presence of a label capable of incorporation into nascent mRNA. Samples are treated with test and control compounds, which will induce mRNA expression in the sample, resulting in incorporation of label. Whole mRNA is isolated and applied to the array such that it hybridizes with the DNAs contained therein. After washing, the amount of hybridization is quantified and a profile is generated. These steps are repeated with various control and test compounds, thereby generating a library of profiles, which can be used to ascertain the relationships relevant to pharmacological efficacy or toxicity.

The matrices used in such profiling, however, need not be limited to those utilizing DNAs. Rather, other nucleic acids, like RNAs and protein nucleic acids (PNAs), as well as the inventive proteins and antibodies corresponding to the inventive proteins may also be employed. Hence, for example, antibodies could form the array and the samples could be treated in order to label nascent proteins. Whole proteins then would be isolated and applied to the antibody matrix. Developing the resulting signal would result in a protein expression profile, which is useful in

essentially the same manner as the nucleic acid profile. A protein matrix could be used, for example, in evaluating antibody responses to pharmaceutical agents in order to eliminate possible cross-reactivity.

5 Moreover, where nucleic acids are used in the matrix, it is often beneficial to use variants (as defined below) of the molecules described hereinin. This can be used to account for genetic variations that are of little or no consequence to the function of the resultant gene product. Hence, they can account
10 for wobble or conservative amino acid variations that do not perturb function, like variations in some of the protein motifs elucidated below. Thus, each position in the matrix can employ multiple nucleic acid probes that account for a series of variants.

15 Expression profiling may also be done, in another embodiment, using two-dimensional protein gels in which the inventive proteins are detected. The resultant profiles can be used in the same way as described.

20 Matrices useful for profiling may be constructed based on different criteria. Of course, the more relevant profiles will take into account expression of most human genes, preferably all of them. In certain situations, however, it is advantageous to look at a smaller subset. For example, if one were concerned about fetal neural toxicity, a fetal brain-specific matrix might
25 be chosen. On the other hand, if one were interested in targeting mammary carcinoma tissue, a corresponding matrix could be used. Thus, matrices may be constructed using all of the sequences available from a tissue-specific library.

* * *

30 The following discussion relates to some of the various functional and structural groupings that would be of interest to the artisan wishing to construct profiling matrices. Of course, the artisan will also recognized that these functional descriptions may find additional applicability in the therapeutic
35 and diagnostic applications discussed below.

Cell Cycle

A proliferating cell must coordinate replication and chromosomal separation to ensure that the genome is replicated

completely, and that a single copy is correctly inherited by each daughter cell. The cell cycle is the coordinated series of events that achieves these aims. Many of the key events are initiated by a family of conserved Serine/threonine protein kinases, the cyclin-dependent kinases (CDKs), that are activated by the cyclin family of proteins (cyclins A-H). In turn, the cyclin-CDK complexes are modulated by other protein kinases or phosphatases, and by binding specific inhibitor proteins. The enormous variety of ways in which CDK activity can be regulated allows the cell to respond to internal signals generated by preceding events in the cell cycle and to external growth signals.

The somatic cell cycle is divided into four phases: DNA replication (S phase) and chromosome separation (M phase) are separated by gap phases (G1 and G2). At specific control points the decision to begin the next stage (DNA synthesis or mitosis) is carefully regulated.

Cdc2, the primary kinase, is especially required for the G1-S transition and S phase. Cdc4 and Cdc6 are involved at the restriction point, where the cell can decide to proliferate or arrest (G1 \leftrightarrow G0) and Cdc7 is a CDK activating kinase (CAK) as well as a subunit of TFIIH.

The Cyclin-CDK complexes are regulated in various ways. One is through phosphorylation by CDK activating kinases (CAK), like the Y15 kinase (Wee1) and dephosphorylation by CDK associated phosphatases (CAP), like Cdc25A a member of the Cdc25 family (Cdc25A, B and C).

An other way of regulation occurs through two classes of CDK inhibitors (CKI), the INK4 proteins p15, p16, p18, and p19, who negatively regulates the cyclin D CDK complexes and second the p21 family with p21, p27, and p57.

The cell cycle is also regulated through ubiquitin-mediated proteolysis involving the destruction of both cyclins and CDK inhibitors by the 26S proteasome, that requires an ubiquitin conjugating enzyme (UBC) and an ubiquitin ligase. The instability is conferred by PEST regions (cyclin D and E) or a ten amino acid

region in the amino terminus (degradation box) in the A- and B-type cyclins.

All these modifications play an important role for the cellular localization, because only the nuclear CDK-cyclin complexes are functional for cell cycle. During G₁ phase of the cell cycle, cyclins A, E and D are synthesized and bind to their cyclin-dependent kinase (CDK) partners. CDK complexes containing cyclins A, E and D₁ are then imported into and concentrated within nuclei. Cdkb- cyclin D₃ has been localized to both cytoplasmic and nuclear compartments, although only the nuclear complex is active. As cells enter S phase, cyclin A and cyclin E complexes remain within the nucleus, whereas cyclin D₁ relocalizes to the cytoplasm for proteolysis at the onset of S phase. Like Cdk2-cyclin A, Cdc2-cyclin A is nuclear and remains so until it is degraded during mitosis. By contrast, as a result of ongoing nuclear import and more rapid re-export, cyclin B₁, which binds to Cdc2 upon synthesis during S phase, is predominantly cytoplasmic. Cdc2-cyclin B₂ is also cytoplasmic, although this might occur through anchoring of the complex to some cytoplasmic constituent. At prophase, phosphorylation of cyclin B₁ promotes accumulation of Cdc2-cyclin B₁ in the nucleus, whereas cyclin B₂ remains in the cytoplasm until nuclear envelope breakdown.

Two crucial regulators of Cdc2-cyclin B-Wee1 and Cdc25C exist and are responsible for the G₂ to M control point. Wee1 is a nuclear protein throughout the cell cycle, whereas Cdc25C binds to 14-3-3 proteins during interphase and remains predominantly cytoplasmic. In some systems Cdc25C, like cyclin B₁, rushes precipitously into the nucleus just before entry into mitosis.

The 110-kDa retinoblastoma (tumor suppressor) protein (RB), a pRB-family member is an important regulator of cell-cycle progression and differentiation. Like the E2F family (E2F1-5) or DP family (DP1-3) of transcription activators, RB suppresses inappropriate proliferation by arresting cells in G₁ by repressing the transcription of genes required for the transition into S phase. Before the cell proceeds into S phase, RB becomes phosphorylated at multiple sites by the cyclin dependent protein

kinases (CDKs) and loses its transcriptional repressing activity. Phosphorylation of RB during late G1 phase results in the dissociation of the E2F-RB repressor complex which allows S-phase specific genes to be transcribed. Cyclin E is the evolutionary
5 conserved target for E2F and interacts together with CDC2 in late G1.

For a proliferating cell it is vital that only undamaged DNA is replicated because if DNA damage is substantial, its replication can lead to chromosome loss or rearrangement. Thus,
10 we find a G1<->S checkpoint in late G1 that requires tumor suppressor p53. A p53-dependent G1 arrest is effected by the cyclin dependent kinase inhibitor p21 through higher expression levels that inhibits almost all cyclin CDK complexes.

The kinase responsible for phosphorylating the unidentified
15 kinetochore component in metaphase may be a member of the MAP kinase family and appears to be the proto oncogene c-MOS, a cytostatic factor (CSF) in meiosis.

Several categories of proteins are coded for by clones of the invention within the overall group of "Cell cycle" and
20 include, among others, the following:

PA26-T2 protein: PA26-T2 is a p53 responsive gene. The protein is predominantly expressed in brain, breast and kidney and represents a novel regulator of cellular growth. Isoforms are
25 differentially induced by genotoxic stress (UV, gamma-irradiation and cytotoxic drugs) in a p53-dependent manner. The p53 tumor antigen is found in increased amounts in a wide variety of transformed cells. The protein is also detectable in many actively proliferating, nontransformed cells, but it is
30 undetectable or present at low levels in resting cells. P53 is postulated to bind as a tetramer to a p53-binding site (PBS) and to activate the expression of adjacent genes that inhibit growth and/or invasion. Deletion or inactivation of one or both p53 alleles reduces the expression of tetramers, resulting in
35 decreased expression of the growth inhibitory genes. This mechanism is found in tumors of several types. (OMIN *191170)
Clones in this category include: amy2_121m2

Cell structure and motility

One of the major differences between prokaryotes and eukaryotes is the ability of the eukaryotic cell to adopt very different shapes dependent on its function during the differentiation process. Animal cells vary from being round to extended cylindric forms like motorneurons or muscle cells. In humans, more than 100 different cell types can be distinguished, each having a characteristic shape. The form of a cell often is closely related to its capacity to move. Some completely differentiated cells like fibroblasts can still change their form actively, thereby migrating. Other cell types serve as motor elements - "macroscopically" like muscle cells or "microscopically" like ciliated epithelia. Such tasks are fulfilled by a big class of proteins; on the one hand responsible for maintenance of cell structure and contacting neighbor cells or the intercellular matrix and on the other hand for cell motility. These topics cannot be regarded separately: The motility apparatus e.g. must be fixed in the cytoskeleton. Three different types of filaments can be distinguished: Actin filaments, tubulin filaments and intermediate filaments, each present in almost all types of cells.

Actin filaments (F-actin) are built up of monomers (G-Actin). In muscle cells, actin, myosin, for both of which several paralogous genes are known, as well as many more proteins are constituents of the contractile apparatus.

The "thin" and "thick filaments" in a muscle cell consist mainly of actin and myosin, respectively.

Several different proteins are responsible for the anchoring of the actin filaments in the Z-disks (e.g. alpha-actinin and desmin) or at the end of the myofibers in the cell membrane.

Troponin I, -C, -T and Tropomyosin - associated with actin - confer the Ca^{++} - dependent triggering of contraction.

Length of the sarcomere is controlled by the giant protein titin.

In smooth muscle, there is no troponin. Contraction activity is controlled by phosphorylation / dephosphorylation of myosin by a specialized kinase instead. Contractile fibers are not organized in sarcomeres.

Apart from contributing to muscle contraction, the actomyosin system is responsible for many other motions at cellular level, e.g. the amoeboid movement of pseudopodia or the fission of cells at the end of mitosis by a contractile ring.

Besides this, actin fibers fulfill structural tasks like maintenance of the shape of stereocilia or microvilli. Here, actin filaments are connected by proteins like fimbrin. But not only specialized structures like the mentioned ones contain actin fibers. There is a network covering the complete cell volume with F-actin as a major constituent. Whereas the actin filaments in the structures mentioned above are relatively stable, this F-actin is highly dynamic. Management of the network structure and turnover is achieved by connecting proteins like alpha-actinin, fimbrin or filipin; turnover is regulated by gelsolin, villin, and different capping- and fragmentation-proteins.

Microtubules are built up of alpha-beta tubulin heterodimers. Turnover of filaments is achieved by building-in and releasing of monomers with different time constant rates at both ends. The resulting cycle is called "treadmilling". Thirteen strings of tubulin duplets build up one subfiber, whereas one fiber contains two or three of those. A complete axoneme consists of 9 radial and 2 central fibers. This "9+2" - structure is the basis both of flagella, their basal bodies and centrioles. In flagella, several additional structures like radial elements exist. Nexin connects the fibers and dyneine is the motor ATPase which shifts the fibers relative to each other. Several genetic diseases like the Cartageneric syndrome are caused by deficiencies of distinct proteins in cilia.

Besides this, microtubules are abundant in all types of cells. They are part of a delivery system for organelles, e.g. in

the golgi apparatus. A further very important system based on microtubules is the mitotic spindle, it is organized by the centrosomes. Besides many other components, the major part of a centrosome are two centrioles which are built up of nine
5 microtubule-triplets. Most remarkably, new centrioles are not synthesized de novo but generated by duplication of old ones.

Cytoplasmic microtubules are associated with many different proteins. Two major classes are known: The MAPs ("microtubule-associated proteins", with molecular masses between 200 and 300
10 kD) and the much smaller tau-Proteins with a MW between 60 and 70 kD. These proteins regulate the treadmill-process and the interaction with other structures in the cell.

Besides actin and myosin the so-called intermediate filaments constitute a third class of filaments. In contrast to
15 the former two groups, they do not participate in motility, nor are they dynamic structures subject to a vivid turnover. The most important ones are neurofilaments (in neurons), keratin filaments (mainly in epithelial cells), and vimentin filaments (in many sorts different cell types).

20 The biological function of both the cytoskeleton as well as contractile apparatus of a cell does not end at the cell membrane. Cells must be embedded in the extracellular matrix, all cells of a muscle must act as one single mechanical unit and epithelia must resist macroscopic mechanical forces. Hence, cell
25 adhesion and the extracellular matrix are closely connected to the cytoskeleton. Vincullin is one of the proteins which serve as an anchor for intracellular fibers (actin). Different types of desmosomes and tight junctions connect neighbor cells with
30 intercellular fibers. On the inside, cytoplasmic plaques connect them to the cytoskeleton. These structures, on the one hand, serve as mechanical elements whereas gap junctions, on the other hand, connect cells metabolically.

The extracellular matrix consists of a network of proteins, glycoproteins and polysaccharides. Different proteins are present
35 in relation to different mechanical demands: Elastin is found in tissues with high elasticity (lungs, heart) whereas collagen,

a more hard-wearing protein, is found in tendons and ligaments. Fibronectin is an extracellular protein highly important for cell adhesion.

Reference: Murray J et al (1992): Cell Motil Cytoskeleton
5 22: 211-223.

Within the overall group of Cell Structure and Motility several categories of proteins are coded for by clones of the invention:

Ankyrins: Ankyrins are peripheral membrane proteins which
10 interconnect integral proteins with the spectrin-based membrane skeleton. Thus these proteins are involved in coupling of cyto skeleton and cell membrane. OMIN reports that Ankyrins have associations (as potentially diagnostic, therapeutic, causative, and/or related, etc...) with the following diseases: 1) Hereditary
15 Spherocytosis (OMIN #182900); 2) Hemolytic Poikilocytic Anemia due to reduced ankyrin binding sites (OMIN 141700); 3) Atypical Elliptocytosis (OMIN 225450); 4) Autosomal recessive spherocystosis (OMIN #270970); 5) Werner Syndrome (OMIN #277700); and 6) Rhesus-unlinked type Elliptocytosis (OMIN #130600).
20 Ankyrin binding glycoprotein proteins mediate Ankyrin effects, especially in neuronal adhesion and prostate tumour vcell transformation: Clones in this category include: amy2_121f19.

Tropomyosins are ubiquitous proteins of 35 to 45 kD associated with the actin filaments of myofibrils and stress
25 fibers. They are involved in cardiomyopathies (OMIN #191030, #191010, #190990, #600317). Clones in this category include: tes3_1bb5.

Differentiation/Development

30 Almost every multicellular organism originates from meiotic cell divisions and the recombination of a paternal and a maternal set of chromosomes. After fertilization of the egg, all cells of a body originate from this one cell. Thus the cells of the developing body are initially genetically alike. But
35 phenotypically they become very different. They are specialized to a certain cell type and arranged in an organized pattern to a certain type of tissue and the whole structure has the well-

defined shape of an organ. All these features are determined by the DNA sequence of the genome, which is reproduced in every cell. Each cell acts on the genetic instructions given to a certain time and at a certain place of development and plays its individual part in the multicellular organism. Cell differentiation may be divided into three general steps: cell cycle exit, apoptosis protection and tissue specific gene expression. These processes are coordinated to provide the final and unique tissue characteristics.

10 An animal cell that has achieved a certain level of development is said to be determined. This differentiation of a cell may be irreversible and in that case the cell may be renewed only by simple duplication. Other cells are renewed by means of stem cells which are immortal (e.g. stem cells of the bone
15 marrow, epidermal stem cells). The genetic control of development is extensively studied in non-vertebrates and vertebrates. The classical animal model is the fruit fly *Drosophila* and the modern model is the transgenic mouse. Animal transgenesis has proven to be useful for physiological as well as
20 physiopathological studies. Besides the approach based on the random integration of a DNA construct in the mouse genome, gene targeting can be achieved using totipotent embryonic stem cells for targeted transgenesis. Transgenic mice are then derived from the embryonic stem cells. This allows the introduction of null
25 mutations in the genome (so-called knock-out) or the control of the transgene expression by the endogeneous regulatory sequence of the gene of interest (so-called knock-in). Mice can be created that express wild-type genes, mutant genes, marker genes or cell lethal genes in a tissue specific manner. These animal
30 models allow to follow changes in tissue and organ development and lead to a better understanding of the cellular function of many genes or to the generation of animal models for human diseases. Fundamental problems in immunology, onset and development of cancer, regulation in fatty acid metabolism,
35 aspects of cardiovascular function, control of the central nervous system development, analysis of reproductive development and function are only some examples of research interests.

The final stage of cell differentiation is growth arrest. In animal tissues with rapid cell turnover terminally differentiated cells undergo programmed cell death. The cells have the ability to kill themselves by activating an intrinsic cell suicide program when they are no longer needed or have become seriously damaged. The execution of this program is termed apoptosis. Apoptosis is of importance for development and homeostasis of animals. The key components of this program have been conserved in evolution from worms (*C. elegans*) to insects (*Drosophila*) to humans. The roles of apoptosis include the sculpting of structures during development, deletion of unneeded cells and tissues, regulation of growth and cell number, and the elimination of abnormal and potentially dangerous cells. In this way apoptosis provides "quality control mechanism" that limits the accumulation of harmful cells, such as virus-infected cells and tumor cells. On the other hand inappropriate apoptosis is associated with a wide variety of diseases, including AIDS, neuro-degenerative disorders and ischemic stroke. Because it is now clear that apoptosis is a result of an active, gene-directed process, it should be eventually possible to manipulate this form of cell death by developing drugs that interact with its recently identified mechanisms of action. Inducers of cell differentiation, cell cycle arrest and apoptosis might be the novel molecular targets for new anticancer agents in addition to the signaling pathways for growth factors and cytokines.

Proteins, factors, receptors and genes of importance in apoptosis:

Proteases:

- Calpain, an intracellular cysteine protease, exact role unknown.
- Caspase-1 to Caspase-11, a family of proteases synthesized as an inactive proenzyme. Targets of the activated enzymes include: poly(ADP-ribose) polymerase, DNA-dependent protein kinase, U1 ribonucleoprotein, nuclear laminins and cytoskeleton components (actin).

- Granzyme B, a serine protease released by cytotoxic T-cells.

Receptors:

5 - CD 95 (synonyms: Fas, AP0-1), a receptor protein of the TNF-receptor family which includes TNF-R1 and TNF-R2 with the common characteristic of a 70 amino acid cytoplasmic domain.

- FADD (synonym: MORT-1), a cytoplasmic protein

- DR-3 (synonym: AP0-3) a member of the TNF-receptor-family

- DR-4 and DR-5

10 Genes:

- ced-3, ced-4 and ced-9 encode the general apoptotic and antiapoptotic program in *Caenorhabditis elegans*. Apaf-3 is the mammalian homologue of ced-3.

15 - Bcl-2 / Bcl-xL / Bax / Bcl-xS / Bak: a large gene family that can either inhibit or promote apoptosis.

- Cytokine response modifier A, a cowpox virus gene whose gene product inhibits caspases.

Others:

20 - Caspase-activated DNase (CAD) and its inhibitor (ICAD), causes DNA fragmentation in the nucleus

- Ceramide, a complex lipid that acts as a second messenger.

- c-Jun N-terminal kinase (JNK) is a proline-directed kinase

- p53 protein, is essential for the induction of apoptosis as a response to chromosomal damage.

25 - RAIDD, a death signal-transducing protein.

- Receptor interacting protein (RIP) is an accessory protein with a death domain and a serine/threonine kinase activity.

- Sphingomyelinase, an enzyme that hydrolyzes the complex lipid sphingomyelin to ceramide.

- Tumor necrosis factor (TNF) is a type -II membrane protein

- TNF-receptor associated factor (TRAF2), is an accessory
5 protein that can bind to both TNF-R1 and TNF-R2.

Within the overall group of Differentiation/Development, several categories of proteins are coded for by clones of the invention:

10 Notch family proteins: Notch family molecules are negative regulators of neuronal differentiation in early brain development. Clones in this category include: amy2_1i24.

15 Testis-specific Y-encoded proteins: The TSPY genes are arranged in clusters on the Y chromosome of many mammalian species. TSPY is believed to function in early spermatogenesis and is a candidate for GBY, the putative gonadoblastoma-inducing gene on the Y. These proteins are involved in early spermatogenesis. Clones in this category include: amy2_7j5.

20 Inflammation-mediating proteins: Inflammation is a basic mechanism responsible for recruiting and activation of immun-competent cells. By various mediators, cells are activated and triggered to differentiate. Hyperactivation of these pathways leads to various disease states: In neuronal tissues, in
25 inflammatory diseases such as experimental autoimmune encephalomyelitis (EAE), neuritis(EAN) and uveitis (EAU) allograft inflammatory factor-1 is produced by macrophages and microglia cells. Clones in this category include: amy2_2b19.

Intracellular transport and trafficking

30 Eukaryotic cells rely for their viability on the partitioning of many basic cellular processes into membrane-bounded organelles. These are the nucleus, endoplasmic reticulum (ER), Golgi apparatus, endosomes, lysosomal compartments, mitochondria and peroxisomes. Most molecules destined for the
35 lysosome, cell surface and outside the cell are routed through

the ER and Golgi, which together with the vesicular intermediates between them, comprise the secretory pathway (Palade 1975). In the ER and Golgi compartments proteins are sorted, modified and often assembled into complexes *en route* to their final

5 destination. Incorrectly assembled proteins are retained in the ER until they fold correctly or are targeted for degradation. Additional proteins are translocated into and function within the luminal spaces of organelles or are secreted. Thus a large proportion of proteins synthesized require targeting to membranes
10 either for insertion into or transport across them. A major purpose of this is growth. The secretory pathway is dependent on an intact cytoskeleton and also closely linked to general metabolism by affecting ribosome biogenesis (Mizuta and Warner, 1994). A huge number of proteins is required for targeting,
15 translocation and sorting of newly synthesized proteins.

The first step in sorting is the recognition of cis-acting targeting or signal sequences that organelle-targeted proteins contain. This is carried out by cytosolic targeting factors and/or receptors on the membrane to which the protein is
20 targeted. In some cases the primary sequences are extremely degenerate, with only the overall character being conserved (hydrophobicity for an ER signal sequence, helical amphiphilicity for mitochondrial targeting sequence (Kaiser et al., 1987; Lemire et al., 1989). Following the targeting step, proteins are either
25 inserted into or transported across the membrane (translocated) through a proteinaceous apparatus (termed the translocon). The translocon include or recruit motors to drive the translocation process in the correct direction (Schatz and Dobberstein, 1996).

Defined intracellular protein transport steps:

- 30 • ER
- targeting to the ER
 - translocation into the lumen of the ER, and, depending on the presence of certain signals in the peptide sequence transport through the golgi complex
- 35 • Mitochondria
- targeting
 - translocation
- Peroxisomes

- The general secretory pathway
 - protein modification, assembly and quality control in the ER
 - vesicle-mediated trafficking
 - vesicle docking and fusion
 - transport through the golgi apparatus and sorting at the trans-golgi
 - transport to the cell surface
 - transport routes to the lysosome

- Endocytosis
- Specialized protein transport routes
- Protein export from the cytoplasm

References: Palade, G (1975) Science 189:347-358; Mizuta et al. (1994) Mol Cell Biol 14: 2493-2502; Kaiser et al. (1987) Science 235: 312-317; Lemire et al. (1989) J Biol Chem 264: 20206-20215; Schatz et al. (1996) Science 271: 1519-1526.

Rab proteins

In eukaryotic cells the compartmentalisation of processes is a prerequisite for a tight regulation of processes and activities. The cells contain a highly dynamic set of membrane compartments that are responsible for packaging, sorting, secreting, and recycling proteins and other molecules. Trafficking between organelles within the secretory pathway occurs as vesicles derived from a donor compartment fuse with specific acceptor membranes, resulting in the directional transfer of cargo molecules. This process is tightly controlled by the Rab/Ypt family of proteins (reviewed by Novick and Zerial, 1997), a branch of the superfamily of small GTPases. Rab proteins regulate a variety of functions, including vesicle translocation and docking at specific fusion sites. Rabs may also play critical roles in higher order processes such as modulating the levels of neurotransmitter release in neurons, a likely mechanism in synaptic plasticity that underlies learning and memory (Geppert and Südhof, 1998).

Small GTPases share a common three-dimensional fold that, in the GTP bound state, can bind a variety of downstream effector proteins. GTP hydrolysis leads to a conformational change in the "switch" regions that renders the GTPase unrecognizable to its

effectors. In this way, by localizing and activating a select set of effectors, a common structural motif is used to control a wide array of distinct cellular processes.

The final steps in membrane fusion are likely to be driven by a set of proteins known as SNAREs. After a vesicle becomes docked, the cytoplasmic domains of VAMP (also termed synaptobrevin) and syntaxin on opposing membranes, in combination with a SNAP-25 molecule, coalesce into an elongated -helical bundle (Poirier et al., 1998; Sutton et al., 1998), which may lead to fusion. Because numerous SNARE isoforms have been identified that localize to distinct membrane compartments, it was originally proposed that the specificity of interaction between the SNARE proteins accounted for the specificity in membrane trafficking. Recent results, however, suggest that SNAREs are not specific in their ability to form complexes in vitro, suggesting that trafficking specificity requires additional factors (Yang et al., 1999). In this regard, Rab proteins are strong candidates for governing the specificity of vesicle trafficking. Like the SNAREs, many isoforms (40) of the Rab family have been identified that localize to specific membrane compartments (reviewed by Novick and Zerial, 1997).

Concomitant with the SNARE cycle, Rab proteins undergo a intricate cycle of membrane and protein interactions. Rabs are posttranslationally modified at C-terminal cysteines by the addition of two geranylgeranyl groups, which mediate membrane association when the Rab is in the GTP-bound state. After guanine nucleotide hydrolysis occurs, the Rab is extracted from the membrane upon forming a complex with a cytosolic GDP-dissociation inhibitor (GDI). This cytosolic intermediate is then recycled onto a newly forming vesicle, most likely through a secondary factor termed a GDI dissociation factor (GDF), which displaces GDI. After the Rab becomes membrane bound, a guanidine nucleotide exchange factor (GEF) promotes release of GDP and the subsequent loading of GTP. In its GTP-bound conformation, the Rab is then free to associate with its specific set of effectors, which can in turn trigger events leading to the eventual fusion of the vesicle with a target membrane. To complete the cycle, perhaps after or concurrent with membrane fusion, a GTPase activating protein (GAP) accelerates nucleotide hydrolysis, switching off

the GTPase. The remaining GDP-bound Rab can then participate in a new round of fusion.

Rab interactions with effectors are likely to regulate vesicle targeting and membrane fusion in three ways. First, a Rab
5 may specifically facilitate vectorial vesicle transport. Vesicles are transported from their site of origin to acceptor compartments likely through associations with cytoskeletal elements and transport motors. A protein has been identified with a domain structure that suggests a connection between the
10 cytoskeleton and the Rabs. This protein, called Rabkinesin-B, contains a kinesin-like ATPase motor domain followed by a coiled-coil stalk region and a RBD that specifically binds Rabb (Echard et al., 1998). An additional link with the cytoskeleton is provided by the Rab effector, Rabphilin-3A. Rabphilin-3A has been
15 shown in vitro to interact with -actinin, an actin-bundling protein, but only when not bound to Rab3A (Kato et al., 1996). These results raise the intriguing possibility that Rab proteins regulate vesicle interactions with the cytoskeleton and thereby play an active role in targeting vesicles to their appropriate
20 destinations.

Second, Rab proteins may regulate membrane trafficking at the vesicle docking step. A number of Rab effectors, including Rabaptin-5, EEA1, Rabphilin-3A, and Rim, may serve as molecular
25 tethers. Each effector protein contains a RBD, followed by a linker region (some having the potential to form elongated coiled-coil structures), and a domain capable of interacting with a second Rab or the target membrane. Rabaptin-5, for example, contains two RBDs, one near the N terminus that specifically recognizes Rab4 and a second near the C terminus that binds Rab5
30 (Vitale et al., 1998). Both Rim, which is localized to the target membrane, and Rabphilin-3A, which is localized to the vesicle, contain N-terminal RBDs and C-terminal Ca^{2+} -binding C2 domains, implicating these effectors in synaptic vesicle localization or docking in response to Ca^{2+} influx (Wang et al.,
35 1997). Tethering effectors may also recognize protein complexes on the acceptor membrane. Sec4p, a yeast Rab3A homolog, interacts with the exocyst (Guo et al., 1999), a complex of seven or more subunits that is assembled at sites of vesicle fusion along the

plasma membrane. The exocyst complex may therefore function as a landmark for Rab/effector-mediated vesicle docking.

Third, once a vesicle has become tethered to its fusion site, Rab proteins may selectively activate the SNARE fusion machinery. The mechanism of this activation is unknown but may involve direct interactions of Rabs or, more likely, their effectors with SNAREs. For example, Hrs-2 is a protein that binds to SNAP-25 and contains a Zn²⁺-finger motif characteristic of Rab-binding proteins such as Rabphilin-3A, Rim, EEA1, and Noc2, suggesting that Hrs-2 may form a physical link between Rabs and SNAREs (Bean et al., 1997). In addition, certain mutations in the syntaxin-binding protein Sly1p, the Sec1p homolog utilized in ER to Golgi trafficking, eliminate the requirement for Ypt1p, a Rab protein that functions at this trafficking step (Dascher et al., 1991). Rabs may therefore regulate SNARE associations through Sec1 family members. In support of this idea, a Rab effector was recently found to interact with a vacuole Rab, a Sec1p homolog, and a SNARE protein (Peterson et al., 1999), which suggests that this effector serves to connect Rab and SNARE function. In this way, Rabs and their effectors may facilitate the correct pairing of SNAREs.

References: Dascher et al. (1991) Mol. Cell. Biol. 11, 872-885; Echard et al. (1998). Science. 279, 580-585; Geppert et al. (1998) Annu. Rev. Neurosci. 21, 75-95; Guo et al. (1999). EMBO J. 18, 1071-1080; Kato et al. (1996) J. Biol. Chem. 271, 31775-31778; Novick et al. (1997) Curr. Opin. Cell Biol. 9, 496-504; Peterson (1999) Curr. Biol. 9, 159-162; Poirier et al. (1998) Nat. Struct. Biol. 5, 765-769; Vitale et al. (1998) EMBO J. 17, 1941-1951; Wang et al. (1997) Nature. 388, 593-598; Yang et al. (1999) J. Biol. Chem. 274, 5649-5653.

Within the overall group of Intracellular Transport and Trafficking several categories of proteins are coded for by clones of the invention.

Vesicular trafficking: Various proteins are involved in trafficking of vesicles inside the cell and for the exocytotic pathway. For example, Sec7 of *Saccharomyces cerevisiae* takes function in vesicular trafficking. Synaptotagmins are essential for Ca²⁺-regulated exocytosis of neurosecretory vesicles. Other proteins such as Dynamin are microtubule-associated force-

producing proteins, which are involved in the production of microtubule bundles. By binding and subsequent hydrolysis of GTP such proteins provide the motor for vesicular transport during endocytosis. Clones in this category include: amy2_14b5,
5 amy_2013 and fkd2_3k1.

Protein sorting: Protein sorting is a process essential for the maintenance of a cell's functionality and structural integrity. Most proteins perform their biological function in special compartments in the cell. The process of sorting is
10 complex and highly regulated. Clones in this category include: mel2_7g14.

Metabolism

This group includes proteins which are involved in the
15 uptake and consumption of nutrients, and enzymes which are part of the biochemical pathways for energy metabolism or which are involved in the supply of building blocks of nucleic acids, proteins (NTPs, dNTPs, amino acids) for DNA/RNA and protein
20 synthesis, and fatty acids (membranes), to allow for the generation of higher order structures. This group constitutes the most important and largest group in prokaryotes and lower eukaryotes. The higher the evolutionary level of an organism is, however, the more other protein classes like 'signal
25 transduction', 'cell cycle' and 'differentiation and development' increase in importance and number of representatives.

Proteins involved in the metabolism of energy and compounds (here: other than nucleic acids or proteins) are usually the products of house keeping genes, they are often constitutively
30 and/or ubiquitously expressed.

Several categories of proteins are coded for by clones of the invention within the overall group of Metabolism:

Fatty acid metabolism: OMIM lists more than 50 diseases caused by pathologic altered fatty acid metabolism. 1-acyl-
35 glycerol-3-phosphate acyltransferase is involved in fatty acid metabolism and is ubiquitously expressed, with a slight predominance in uterus, placenta and foreskin. Clones in this category include: amy2_2c22

Repair and surveillance of protein damage: Several classes of protein are involved in repair and surveillance of protein damage. L-isoaspartyl methyltransferase (Pimt), as an example, is a highly conserved enzyme utilising S-adenosylmethionine (AdoMet) to methylate aspartate residues of proteins damaged by age-related isomerisation and deamidation. Clones in this category include: fbr2_78i21.

Nucleic acid management

The genetic information is stored in the form of nucleic acids in all organisms. Two kinds of nucleic acids exist, DNA and RNA. Whereas the more stable DNA in most organisms constitutes the storage form of the genetic information, the labile RNA and in particular mRNA is an intermediate used for the temporal expression of specific genes.

In eukaryotes, DNA is usually a double stranded linear molecule consisting of two antiparallel strands and made up of a deoxyribose, a phosphorus backbone and the four bases A, C, G, and T. The DNA of some organisms has a ring structure. The structure of DNA was unraveled years ago by Watson and Crick. DNA is directional molecule determined by the C-atoms of the sugar.

The most important processes dealing with nucleic acids are:

- replication (e.g. DNA polymerases, Telomerase)
- transcription (RNA polymerases)
- RNA processing (maturation - splicing and degradation)
- in addition, enzymes and proteins exist which require a nucleic acid (mostly RNA) in the active center to be functional (ribozymes - e.g. RNase, Ribosomal proteins)

The DNA of a cell is replicated in the S-phase of the cell cycle. Several enzymes carry out the task of doubling this nucleic acid. As all steps of the cell cycle, also the process of replication is tightly regulated. The enzyme DNA polymerase and several other proteins are involved in this process. Whereas many prokaryotes do have only one origin of replication (i.e., the starting point of the replication cycle), in eukaryotic DNAs (chromosomes) multiple such start points exist. The switch from the synthesis (S) phase to the subsequent G2 or M phases of the cell cycle are dependent on the completion of the replication.

This makes clear, that a number of proteins are involved in the replication itself as well as in the control of the process. Since most eukaryotic chromosomes are linear structures, additional proteins and enzymes are necessary to make sure that the structure is maintained through successive generations. This includes those proteins necessary to build the three dimensional structure of chromosomes (e.g. histones) and the structural network of the nucleus and nucleolus (including the defined localization of transcriptionally active genes in the vicinity of nucleoli) but also such enzymes as telomerase which guarantees the integrity of the chromosomal ends.

The expression of genes is usually performed in two steps. First a messenger RNA (mRNA) is produced (transcribed) in one to many copies and second this mRNA is translated into the protein product. The regulation of transcription is discussed under the separate heading 'transcription factors', but also the classes 'signal transduction', 'development', 'cell cycle' and others are affected as the expression of certain genes determines the fate of a cell or organism.

The primary transcript (hnRNA - heterogeneous nuclear RNA) is a single stranded one-to-one copy of the gene as it is located on the chromosome. Before a protein can be translated, already during transcription the process of maturation is initiated. Firstly, a 5' cap structure is enzymatically and covalently added to the RNA, blocking the 5' end of the RNA. Second, when the RNA polymerase has terminated polymerization, the enzyme poly A polymerase adds varying numbers of adenine residues to the 3' end of the transcript. This enzyme recognizes the sequence AAUAAA or AUUAAA (+ some minor variations), cuts the RNA 10 - 30 nucleotides downstream and adds the A residues. The size of the poly A sequence affects the stability of the RNA. Finally, in the process of splicing, the introns present on the genomic level and also present in the hnRNA are spliced out by a multi-protein complex consisting of several proteins and RNAs. The finally matured mRNA is exported to the cytoplasm where it is translated with help of the ribozymes.

The half life of RNA is usually much shorter than that of DNA. Usually, the mRNA is degraded shortly after synthesis, to guarantee a very defined window of expression of a given gene.

This regulation is necessary to specifically maintain or change the set of proteins present at any time in a cell. Specific regions in the 3'UTR (untranslated region) determine the stability of the mRNA in the cytoplasm before it is degraded by RNases, enzymes consisting both of protein and RNA.

References: Watson and Crick (1953) Nature 171: 737-738.

Several categories of proteins are coded for by clones of the invention within the overall group of "Nucleic acid management" and include, among others, the following:

Proteins induced by DNA-Damage: There are several distinct pathways responsible for repair of DNA. Nucleotide excision repair is the most versatile DNA repair pathway and is the main defense of mammalian cells against UV-induced DNA damage. Defects in proteins involved in this pathway can lead to inherited disorders (such as xeroderma pigmentosum OMIN *278700, *278720, *278740 and *194400; Cockayne's syndrome OMIN *216400 and trichothiodystrophy OMIN #601675). Study of UV-sensitive yeast RAD mutants has greatly aided this process and has revealed strong conservation of the components of nucleotide excision repair in eukaryotes. Clones in this category include: amy2_11n4 and tes3_10i16.

Proteins involved in Loading of transfer RNAs: transfer RNAs must be coupled to an amino acid, which then is transported to the peptideyl-transferase centre of the ribosome. Clones in this category include: fbr2_78c12.

Cytosolic ribosomal proteins: Several proteins are part of the eukaryotic ribosomal peptidyl transferase center or modulate the activity of this centre. Such proteins can find application in modulation of ribosome assembly, maintenance and activity. Clones in this category include: amy21i1

Histones: Histones are DNA-binding protein responsible not only for DNA structure and folding and packing, but also are discussed to be involved in activation and silencing of large chromosomal regions. Clones in this category include: tes3_31a10.

mRNA-binding proteins: mRNA-binding are involved in regulation of mRNA folding, translation and stability. For example, the VILIP protein binds specifically to the

3'untranslated region of the neurotrophin receptor mRNA. Clones in this group include amy2_2g12.

Signal transduction

Cells in higher order organisms need to continuously communicate with its environment especially with other cells of the same organism in order to maintain the function and specialization of the whole system these cells are part of. This important task of communication is performed with help of cell-surface receptors which receive and transmit signals from outside into the cell.

G-proteins

The largest known family of cell-surface receptors is that of the G-protein-coupled receptors, which mediate the transmission of diverse stimuli such as neurotransmitters, glycopeptides, hormones, peptides, odorant molecules, and photons. The functional unit of these receptors is composed of the receptor molecule itself (GPCR) which is anchored in the cytoplasmic membrane with seven membrane spanning domains, the heterotrimeric G-protein which is composed of α and β -subunits (G_α and G_β), and the effectors that interact with G_α and / or G_β . In particular, the dissociated G_α and G_β can regulate the activities of a number of effector molecules such as adenylate cyclases, phospholipase C isoforms, ion channels, and tyrosine kinases, resulting in a variety of cellular functions. The process of signal transduction must be tightly regulated and reversible in order to avoid overstimulation, to achieve signal termination, and render the receptor responsive to subsequent stimuli [Iacovelly L. et al., (1999) *FASEB J.* 13, 1-8, Hamm, H.E. (1998) *J. Biol. Chem.* 273, 669-672].

G-proteins are GTPases that, upon binding of GTP change their conformation which in return unmasks structural motives, in particular the so called effector loop, which can mediate the interactions to target proteins, or effectors, for the GTPases. This ability enables the GTPases to cycle between active, GTP-bound and inactive, GDP bound conformations and in the process to function as molecular traffic lights in a multitude of signal transduction pathways. The most important of these signal transduction pathways that are regulated with help of G-proteins

are that of the phospholipase C / protein kinase C and that of the adenylate cyclase / protein kinase A.

The cycling of GTPases is tightly regulated by three main classes of proteins: The exchange of hydrolyzed GDP for a fresh GTP is facilitated by guanosine nucleotide exchange factors (GEFs), the hydrolysis of GTP to GDP is sped up by GTPase-activating proteins (GAPs), and the dissociation of GDP from the GTPases is inhibited by GDP dissociation inhibitors (GDIs) [Tapon and Hall (1997) *Curr.Opin. Cell. Biol.* 9, 86-92, Van Aelst and D-Souza-Schorey (1997) *Genes Dev.* 11, 2295-2322].

S0C-family

A conserved motif that was originally identified in proteins that negatively regulate the signaling action of cytokines was termed S0CS box, the Suppressor Of Cytokine Signaling. Based on homology, five distinct structural protein classes have been identified since that carry this motif. The function of most of these proteins is presently not known. Common to the proteins is only the S0CS box which is located near the C-terminus of the respective peptides. Recently, the S0CS box has been demonstrated to induce binding of proteins to elongins B and C which could target the proteins (and bound substrates) to the proteasomal protein degradation pathway (Kamura, T. et al. (1998) *Genes Dev.* 12, 3872-3881; Zhang, J.-G. et al. (1999) *Proc. Natl. Acad. Sci. USA* 96, 2071-2076).

The class where the S0CS box was originally described contains several members (S0CS-1-S0CS-7 and CIS). In addition to the S0CS box, these proteins also contain a SH2 (Src-homology 2) domain and a variable N-terminus. These S0CS proteins appear to form part of a classical negative feedback loop that regulates cytokine signal transduction. Upon cytokine stimulation, expression of S0CS proteins is rapidly induced and the proteins inhibit further cytokine action. The mode of action of the S0CS proteins is variable. While S0CS-1 binds and inhibits the JAK (Janus kinases) family of cytoplasmic protein kinases [Narahzaki M. et al. (1998) *Proc. Natl. Acad. Sci. USA* 95, 13130-13134, Nicholson, S.E. et al. (1999) *EMBO. J.* 18, 375-385], CIS appears to act by competing with signaling molecules such as the STATs (Transducers and Activators of Transcription) family for binding

to phosphorylated receptor cytoplasmic domains [Yoshimura, A. et al. (1995) *EMBO J.* 14, 2816-2826; Matsumoto, A. et al. (1997) *Blood* 89, 3148-3154].

A second class of SOCS box protein contains additionally WD-40 repeats which were initially identified in the mouse WSB-1 and -2 proteins. The functions of WD-40 proteins are not completely understood but seem to be rather divergent. In Cdc4p the WD-40 repeats probably are necessary for binding the substrate for Cdc34p [Mathias, N. et al. (1999) *Mol. Cell Biol.* 19, 1759-1767].

Cdc4p is a component of a ubiquitin ligase that tethers the ubiquitin-conjugating enzyme Cdc34p to its substrates. The posttranslational modification of a protein by ubiquitin usually results in rapid degradation of the ubiquitinated protein by the proteasome. The transfer of ubiquitin to substrate is a multistep process where WD-40 repeats might play an important function.

Other WD-40 containing proteins (e.g. the retino blastoma binding protein RbAp48) have been shown to bind metal ions (Zinc) and that this metal binding might mediate and/or regulate protein-protein interactions which are functionally important in chromatin metabolism [Kenzior, A.L. and Folk, W.R. (1998) *FEBS Lett.* 440, 425-429]. These proteins are involved in the RAS-cAMP pathway that regulates cellular growth [Ach R.A. et al. (1997) *Plant Cell* 9, 1595-1606].

The SPRY domain has been identified in pyrin or marenostrin, a protein which is mutated in patients with Mediterranean fever and which is similar to the butyrophilin family. While butyrophilins seem to be involved in the lactation process in mammals, the function pyrin is unknown. Three proteins (SSB-1 to -3) have been identified to contain both SPRY and SOCS box motifs. The function of these proteins is also not known.

Ankyrin repeat containing proteins share a 33-residue repeating motif, an L-shaped structure with protruding -hairpin tips which mediate specific macromolecular interactions with cytoskeletal, membrane, and regulatory proteins. These proteins play fundamental roles in diverse biological activities including growth and development, intracellular protein trafficking, the establishment and maintenance of cellular polarity, cell adhesion signal transduction, and mRNA transcription. Three proteins that

contain ankyrin repeats (ASB-1 to -3) have been identified to contain a C-terminal S0CS box additionally to the ankyrin repeats. The function of these proteins or the individual domains remains to be discovered [Hilton, D.J. et al. (1998) *Proc. Natl.*

5 *Acad. Sci. USA* 95, 114-119].

A few small GTPases (RAR and RAR like) do also contain a S0CS box. GTPases are involved in signal transduction during cellular communication. The function of the S0CS box in this type of proteins is currently unclear [Hilton, D.J. et al. (1998)

10 *Proc. Natl. Acad. Sci. USA* 95, 114-119].

Ca²⁺ as second messenger

The bivalent cation Ca²⁺ is, besides cAMP, one of the two major second messengers in eukaryotic cells. Its intracellular concentration is tightly regulated and usually kept very low compared to the cell's environment. Ca²⁺ binding proteins and transporters (Gap junction, Voltage-gated, second messenger-gated) help to sequester huge amounts of the ion in various organelles from where Ca²⁺ can be released upon extracellular stimuli. E.g. the contraction of the muscle is dependent on the presence of Ca²⁺ ions which are readily transported back into the organelles in order for the muscle to relax. In signal transduction, Ca²⁺ functions as a second messenger that activates Ca²⁺ dependent processes through the activation of Ca²⁺/calmodulin dependent protein kinases (CaM kinases) which are the major effector molecules of Ca²⁺. In the signaling cascades, the CaM dependent kinases activate phospholipases (e.g. phospholipase C) that in return activate other protein kinases such as protein kinase C.

cAMP

The cyclic AMP is produced by the enzyme adenylate cyclase in response to extracellular signals. Certain G-proteins stimulate the activity of adenylate cyclase which converts ATP to cAMP and PPi. Two molecules of cAMP bind to each of two regulatory subunits of cAMP dependent protein kinase which in turn dissociate from the two catalytic subunits of the heterotetramer R₂C₂. Upon release of the C-subunits, they become active and phosphorylate substrate proteins at Ser and Thr residues. The process leading from binding of extracellular

molecules to their receptors, the transmission of the stimuli into the cell, the activation of adenylate cyclase and the subsequent activation of cAMP dependent protein kinase is one of two major signal transduction pathways in eukaryotic cells. Since
5 the phosphorylation of proteins is a posttranslational modification of proteins, the kinases are described in the class "signal transduction."

SARA

Members of the transforming growth factor β (TGF β)
10 superfamily signal through a family of cell-surface transmembrane serine/threonine kinases, known as type I and type II receptors (Heldin et al., 1997 ; Attisano and Wrana, 1998 ; Kretzschmar and Massagué, 1998). Ligand induces formation of heteromeric complexes of these receptors, and signaling is initiated when
15 receptor I is phosphorylated and activated by the constitutively active kinase of receptor II (Wrana et al., 1994). The activated type I receptor kinase then propagates the signal to a family of intracellular signaling mediators known as Smads (contraction of the C.elegans Sma and Drosophila Mad genes which were the first
20 identified members of this class of signaling effectors).

Three classes of Smads with distinct functions have been defined: the receptor-regulated Smads, which include Smad1, 2, 3, 5, and 8; the common mediator Smad, Smad4; and the antagonistic Smads, which include Smad6 and 7 (Heldin et al., 1997; Attisano and Wrana, 1998 ; Kretzschmar and Massagué, 1998). Receptor-regulated Smads (R-Smads) act as direct substrates of specific
25 type I receptors, and the proteins are phosphorylated on the last two serines at the carboxyl terminus within a highly conserved SSXS motif (Macías-Silva et al., 1996 ; Abdollah et al., 1997 ; Kretzschmar et al., 1997 ; Liu et al., 1997b ; Souchelnytskyi et al., 1997). Regulation of R-Smads by the receptor kinase provides an important level of specificity in this system. Thus, Smad2 and Smad3 are substrates of TGF β or activin receptors and mediate signaling by these ligands (Macías-Silva et al., 1996 ;
30 Liu et al., 1997b ; Nakao et al., 1997), whereas Smad1, 5, and 8 are targets of BMP receptors and propagate BMP signals (Hoodless et al., 1996 ; Chen et al., 1997b ; Kretzschmar et al., 1997 ; Nishimura et al., 1998). Once phosphorylated, R-Smads associate with the common Smad, Smad4 (Lagna et al., 1996 ; Zhang et al.,

1997), and mediate nuclear translocation of the heteromeric complex. In the nucleus, Smad complexes then activate specific genes through cooperative interactions with DNA and other DNA-binding proteins such as FAST1, FAST2, and Fos/Jun (Chen et al., 1996 ; Chen et al., 1997a ; Liu et al., 1997a ; Labbé et al., 1998 ; Zhang et al., 1998 ; Zhou et al., 1998). In contrast to R-Smads and Smad4, the antagonistic Smads, Smad6 and 7, appear to function by blocking ligand-dependent signaling (reviewed in Heldin et al., 1997).

Phosphorylation of R-Smads by the type I receptor is essential for activating the TGF β signaling pathway (Heldin et al., 1997 ; Attisano and Wrana, 1998 ; Kretzschmar and Massagué, 1998). However, little is known of how Smad interaction with receptors is controlled. A novel Smad2/Smad3 interacting protein has been described (Tsukazaki T. et al., 1998) that contains a double zinc finger, or FYVE domain, and which has been called SARA (Smad anchor for receptor activation). The SARA motif recruits Smad2 into distinct subcellular domains and co-localizes and interacts with TGF β receptors. TGF β signaling induces dissociation of Smad2 from SARA with concomitant formation of Smad2/Smad4 complexes and nuclear translocation. Moreover, deletion of the FYVE domain in SARA causes mislocalization of Smad2 and inhibits TGF β -dependent transcriptional responses. Thus, SARA defines a component of TGF β signaling that functions to recruit Smad2 to the receptor by controlling the subcellular localization of Smad.

References: Abdollah et al. (1997) J. Biol. Chem. 272, 27678-27685; Attisano et al. (1998) Curr. Opin. Cell Biol. 10, 188-194; Chen et al. (1996) Nature 383, 691-696; Chen et al. (1997a) Nature 389, 85-89; Chen et al. (1997b) Proc. Natl. Acad. Sci. USA 94, 12938-12943; Heldin et al. (1997) Nature 390, 465-471; Hoodless et al. (1996) Cell 85, 489-500; Kretzschmar et al. (1998) Curr. Opin. Genet. Dev. 8, 103-111; Kretzschmar et al. (1997) Genes Dev. 11, 984-995; Labbé et al. (1998) Mol. Cell 2, 109-120; Lagna et al. (1996) Nature 383, 832-836; Liu et al. (1997a) Genes Dev. 11, 3157-3167; Liu et al. (1997b) Proc. Natl. Acad. Sci. USA 94, 10669-10764; Macías-Silva et al. (1996) Cell 87, 1215-1224; Nakao et al. (1997) EMBO J. 16, 5353-5362; Nishimura et al. (1998) J. Biol. Chem.

- 273, 1872-1879; Souchelnytskyi et al. (1997) J. Biol. Chem.
272, 28107-28115; Tsukazaki et al. (1998) Cell 95, 779-791;
Wrana et al. (1994) Nature 370, 341-347; Zhang et al. (1997)
Curr. Biol. 7, 270-276; Zhang et al. (1998) Nature 394, 909-
5 913; Zhou et al. (1998) Mol. Cell 2, 121-127.

Calcium

The bivalent cation Ca^{2+} is, along with cAMP, one of the two major second messengers in eukaryotic cells. Its intracellular
10 concentration is tightly regulated and usually kept very low compared to the cell's environment. Ca^{2+} binding proteins and transporters (Gap junction, Voltage-gated, second messenger-gated) help to sequester huge amounts of the ion in various organelles from where Ca^{2+} can be released upon extracellular
15 stimuli. E.g. the contraction of the muscle is dependent on the presence of Ca^{2+} ions which are readily transported back into the organelles in order for the muscle to relax. In signal transduction, Ca^{2+} functions as a second messenger that activates Ca^{2+} dependent processes through the activation of Ca^{2+} /calmodulin
20 dependent protein kinases (CaM kinases) which are the major effector molecules of Ca^{2+} . In the signaling cascades, the CaM dependent kinases activate phospholipases (e.g. phospholipase C) that in return activate other protein kinases such as protein kinase C.

25 Rab proteins

In eukaryotic cells the compartmentalization of processes is a prerequisite for a tight regulation of processes and activities. The cells contain a highly dynamic set of membrane compartments that are responsible for packaging, sorting,
30 secreting, and recycling proteins and other molecules. Trafficking between organelles within the secretory pathway occurs as vesicles derived from a donor compartment fuse with specific acceptor membranes, resulting in the directional transfer of cargo molecules. This process is tightly controlled
35 by the Rab/Ypt family of proteins (reviewed by Novick and Zerial, 1997), a branch of the superfamily of small GTPases. Rab proteins regulate a variety of functions, including vesicle translocation and docking at specific fusion sites. Rabs may also play critical roles in higher order processes such as modulating

the levels of neurotransmitter release in neurons, a likely mechanism in synaptic plasticity that underlies learning and memory (Geppert and Südhof, 1998).

5 Small GTPases share a common three-dimensional fold that, in the GTP bound state, can bind a variety of downstream effector proteins. GTP hydrolysis leads to a conformational change in the "switch" regions that renders the GTPase unrecognizable to its effectors. In this way, by localizing and activating a select set of effectors, a common structural motif is used to control a wide
10 array of distinct cellular processes.

The final steps in membrane fusion are likely to be driven by a set of proteins known as SNAREs. After a vesicle becomes docked, the cytoplasmic domains of VAMP (also termed synaptobrevin) and syntaxin on opposing membranes, in combination
15 with a SNAP-25 molecule, coalesce into an elongated -helical bundle (Poirier et al., 1998 ; Sutton et al., 1998), which may lead to fusion. Because numerous SNARE isoforms have been identified that localize to distinct membrane compartments, it was originally proposed that the specificity of interaction
20 between the SNARE proteins accounted for the specificity in membrane trafficking. Recent results, however, suggest that SNAREs are not specific in their ability to form complexes in vitro, suggesting that trafficking specificity requires additional factors (Yang et al., 1999). In this regard, Rab
25 proteins are strong candidates for governing the specificity of vesicle trafficking. Like the SNAREs, many isoforms (40) of the Rab family have been identified that localize to specific membrane compartments (reviewed by Novick and Zerial, 1997).

Concomitant with the SNARE cycle, Rab proteins undergo a
30 intricate cycle of membrane and protein interactions. Rabs are posttranslationally modified at C-terminal cysteines by the addition of two geranylgeranyl groups, which mediate membrane association when the Rab is in the GTP-bound state. After guanine nucleotide hydrolysis occurs, the Rab is extracted from the
35 membrane upon forming a complex with a cytosolic GDP-dissociation inhibitor (GDI). This cytosolic intermediate is then recycled onto a newly forming vesicle, most likely through a secondary factor termed a GDI dissociation factor (GDF), which displaces GDI. After the Rab becomes membrane bound, a guanidine nucleotide

exchange factor (GEF) promotes release of GDP and the subsequent loading of GTP. In its GTP-bound conformation, the Rab is then free to associate with its specific set of effectors, which can in turn trigger events leading to the eventual fusion of the vesicle with a target membrane. To complete the cycle, perhaps after or concurrent with membrane fusion, a GTPase activating protein (GAP) accelerates nucleotide hydrolysis, switching off the GTPase. The remaining GDP-bound Rab can then participate in a new round of fusion.

Rab interactions with effectors are likely to regulate vesicle targeting and membrane fusion in three ways. First, a Rab may specifically facilitate vectorial vesicle transport. Vesicles are transported from their site of origin to acceptor compartments likely through associations with cytoskeletal elements and transport motors. A protein has been identified with a domain structure that suggests a connection between the cytoskeleton and the Rabs. This protein, called Rabkinesin-b, contains a kinesin-like ATPase motor domain followed by a coiled-coil stalk region and a RBD that specifically binds Rab6 (Echard et al., 1998). An additional link with the cytoskeleton is provided by the Rab effector, Rabphilin-3A. Rabphilin-3A has been shown in vitro to interact with γ -actinin, an actin-bundling protein, but only when not bound to Rab3A (Kato et al., 1996). These results raise the intriguing possibility that Rab proteins regulate vesicle interactions with the cytoskeleton and thereby play an active role in targeting vesicles to their appropriate destinations.

Second, Rab proteins may regulate membrane trafficking at the vesicle docking step. A number of Rab effectors, including Rabaptin-5, EEA1, Rabphilin-3A, and Rim, may serve as molecular tethers. Each effector protein contains a RBD, followed by a linker region (some having the potential to form elongated coiled-coil structures), and a domain capable of interacting with a second Rab or the target membrane. Rabaptin-5, for example, contains two RBDs, one near the N terminus that specifically recognizes Rab4 and a second near the C terminus that binds Rab5 (Vitale et al., 1998). Both Rim, which is localized to the target membrane, and Rabphilin-3A, which is localized to the vesicle, contain N-terminal RBDs and C-terminal Ca^{2+} -binding C2

domains, implicating these effectors in synaptic vesicle localization or docking in response to Ca^{2+} influx (Wang et al., 1997). Tethering effectors may also recognize protein complexes on the acceptor membrane. Sec4p, a yeast Rab3A homolog, interacts with the exocyst (Guo et al., 1999), a complex of seven or more subunits that is assembled at sites of vesicle fusion along the plasma membrane. The exocyst complex may therefore function as a landmark for Rab/effector-mediated vesicle docking.

Third, once a vesicle has become tethered to its fusion site, Rab proteins may selectively activate the SNARE fusion machinery. The mechanism of this activation is unknown but may involve direct interactions of Rabs or, more likely, their effectors with SNAREs. For example, Hrs-2 is a protein that binds to SNAP-25 and contains a Zn^{2+} -finger motif characteristic of Rab-binding proteins such as Rabphilin-3A, Rim, EEA1, and Noc2, suggesting that Hrs-2 may form a physical link between Rabs and SNAREs (Bean et al., 1997). In addition, certain mutations in the syntaxin-binding protein Slylp, the Sec1p homolog utilized in ER to Golgi trafficking, eliminate the requirement for Ypt1p, a Rab protein that functions at this trafficking step (Dascher et al., 1991). Rabs may therefore regulate SNARE associations through Sec1 family members. In support of this idea, a Rab effector was recently found to interact with a vacuole Rab, a Sec1p homolog, and a SNARE protein (Peterson et al., 1999), which suggests that this effector serves to connect Rab and SNARE function. In this way, Rabs and their effectors may facilitate the correct pairing of SNAREs.

References: Dascher et al. (1991). *Mol. Cell. Biol.* 11, 872-885; Echard et al. (1998). *Science*. 279, 580-585; Geppert et al. (1998). *Annu. Rev. Neurosci.* 21, 75-95; Guo et al. (1999). *EMBO J.* 18, 1071-1080; Kato et al. (1996). *J. Biol. Chem.* 271, 31775-31778; Novick et al. (1997). *Curr. Opin. Cell Biol.* 9, 496-504; Peterson et al. (1999). *Curr. Biol.* 9, 159-162; Poirier et al. (1998). *Nat. Struct. Biol.* 5, 765-769; Vitale et al. (1998). *EMBO J.* 17, 1941-1951; Wang et al. (1997). *Nature*. 388, 593-598; Yang et al. (1999). *J. Biol. Chem.* 274, 5649-5653.

Kinases

Reversible posttranslational modifications of proteins are major means of regulating cellular activities. Among the various modifications that are carried out by the cells, the addition of phosphoryl groups to Ser/Thr or Tyr residues is the most important and widely used. The phosphorylation of proteins is accomplished by protein kinases, while the reverse reaction, the removal of phosphoryl groups, is carried out by phosphatases. Kinases / Phosphatases regulate key positions e.g. in the processes of cell proliferation, differentiation and communication/signaling. These processes must be tightly regulated in order to maintain a steady state level of cellular fate. Mis-regulation of kinase activities (or that of phosphatases) is made responsible for a multitude of disease processes such as oncogenesis, inflammatory processes, arteriosclerosis, and psoriasis.

Protein kinases constitute the largest protein family that is currently known. Several hundred kinases have been identified already. Classically, kinases are subdivided into two classes based on the amino acid residues in their substrates that are phosphorylated by the particular enzymes. The kinases specifically add phosphoryl groups from adenosine triphosphate (ATP) or, less frequently, guanosine triphosphate (GTP), either to serine and/or threonine or to tyrosine residues of substrate proteins. An estimated 1,000 to 10,000 proteins present in a typical mammalian cell are believed to be regulated also by the action of protein kinases.

Protein kinases are frequently integral parts of signaling cascades that transmit extracellular stimuli (e.g. hormones, neurotransmitters, growth- or differentiation factors) into the cell and result in various responses by the cells. The kinases play key roles in these cascades as they constitute a sort of 'molecular switches' turning on or off the activities of other enzymes and proteins, e.g. metabolic, regulatory, channels and pumps, receptors, cytoskeletal, transcription factors.

The regulation of kinase activities is accomplished by various means:

The best characterized example for the regulation via regulatory subunits is the cAMP-dependent protein kinase (PKA) which is also a prototype for second messenger activated protein

kinases. This enzyme consists of a heterotetramer of two catalytic (C) and two regulatory (R) subunits. Upon binding of two molecules of second messenger (cAMP) in each R subunit, the catalytic subunits are released and active. Both of the catalytic and the regulatory subunits several isoforms exist. The combination of catalytic and regulatory subunits determines the localization of the holoenzyme and also the substrate spectrum that is available for phosphorylation. The consensus pattern necessary to be present in the substrate for PKA action is RRXS/T where X can be any amino acid.

The casein kinase II comprises another examples for holoenzymes that consist of catalytic and regulatory subunits. Other kinases that are activated by second messengers are cGMP-dependent protein kinase and Protein kinase C (PKC) which is activated by diacylglycerol, which in-turn is produced by phospholipases by cleavage of phosphatidylcholine.

Receptor kinases usually consists of an extracellular domain which can bind effector molecules (e.g. growth factors and hormones) and transfer the stimulus to the intracellular domain of these proteins which usually is a protein tyrosine kinase. Other tyrosine kinases lack an extracellular domain but are associated with receptors which transfer the signal after effector binding by activating the associated protein kinase enzyme (e.g. Src kinase family; Src, Blk, Fgr, Fyn, Lck Lyn, Yes and Janus kinase family; Jak1-3, Tyk2).

Dysfunction of kinases, e.g. caused by non-functioning regulation, can be the cause of inflammatory diseases and uncontrolled proliferation. v-Src which is a truncated version of the C-Src protooncogene tyrosine kinase is a classical example for this process as v-Src does not contain the regulatory domain of the cellular gene and is thus constitutively active.

Several categories of proteins are coded for by clones of the invention within the overall group of "Signal transduction" and include, among others, the following:

Discs-large family: In *Drosophila* more than 50 genes are described in which mutation leads to loss of cell proliferation control indicating that they are tumor suppressor genes. Most of

these genes have mammalian homologs. The *Drosophila* 'discs large' tumor suppressor protein, Dlg, is the prototype of a family of proteins termed MAGUKs (membrane-associated guanylate kinase homologs). MAGUKs are localized at the membrane-cytoskeleton interface, usually at cell-cell junction, where they appear to have both structural and signaling roles. They contain several distinct domains, including a modified guanylate kinase domain, an SH3 motif, and 1 or 3 copies of the DHR (GLGF/PDZ) domain. Recessive lethal mutations in the 'discs large' tumor suppressor gene interfere with the formation of septate junctions (thought to be the arthropod equivalent of tight junctions) between epithelial cells, and they also cause neoplastic overgrowth of imaginal discs, suggesting a role for cell junctions in proliferation control. These proteins can find application in modulating/blocking the guanylate cyclase-pathway. Clones in this category include: amy2_12d7.

Proteins with a WW Domain: Proteins that contain a WW domain which has been originally described as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown to bind proteins with particular proline-motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. This domain is frequently associated with other domains typical for proteins in signal transduction processes. Examples of proteins containing the WW domain are Dystrophin, Utrophin, vertebrate YAP protein (binds the SH3 domain of the Yes oncoprotein), murine NEDD-4 (embryonic development and differentiation of the central nervous system), IQGAP (human GTPase activating protein acting on ras). Therefore these proteins should be involved in intracellular signal transduction. Diseases associated (as potentially diagnostic, therapeutic, causative, and/or related, etc...) with these proteins include as reported by OMIM 1) Muscular Dystrophy, Pseudohypertrophic Progressive Duchenne and Becker Types (OMIM *310200). Clones in this category include: tes3_11d21.

Ion-Transporters: For signalling stringent control of ion fluxes over biological membranes is of the essence. Several trans-membrane ion-channel-proteins key elements of signal transduction pathways. Clones in this category include: amy2_10p7
5 and amy2_2f18.

RING-finger proteins: A Zinc finger motif of the C3HC4 type (the so-called RING finger domain) is involved in mediating protein-protein interactions. Proteins containing a RING-finger are: mammalian V(D)J recombination activating protein (RAG1),
10 mouse rpt-1, human rfp, human 52 Kd Ro/SS-A protein and others. The family of RING finger proteins contains a number of oncogenes. For example PML, a probable transcription factor, BRCA1, the mammalian cbl- and bmi-1 proto-oncogenes. Clones in this category include: amy2_10h17.

15 Phosphatases: Proper targeting of PTPs is essential for many cellular signalling events including antigen induced proliferative responses of B and T cells. The physiological significance of PTPs is further unveiled through mice gene knockout studies and human genome sequencing and mapping
20 projects. Several PTPs are shown to be critical in the pathogenesis of human diseases, as shown by over 290 entries in OMIM. Clones in this category include: tes3_31j20.

Phosphoproteins: Some paraneoplastic syndromes affecting the nervous system are associated with antibodies that react with
25 neuronal proteins and the causal tumor (onconeurological antigens). Several of these antibodies are markers of specific neurologic syndromes associated with distinct types of cancer. One of the antigens recognised by such antibodies is Ma-1, the neuron- and testis-specific protein 1. The expression of Mal mRNA is highly
30 restricted to the brain and testis. Subsequent analysis suggested that Mal is likely to be a phosphoprotein (see OMIM *604010). Clones in this category include: tes3_5k22.

Transmembrane proteins

Membrane region prediction was effected using the ALOM2
35 software (Klein et al., 1985; version 2 by K. Nakai). Similar to

many other methods, the Kyte & Doolittle (1982) amino acid hydrophobicity scale is used in ALOM2 as the primary variable for classifying sequences in terms of their localization. High prediction accuracy is achieved through the system of intelligent decision rules and the utilization of a carefully selected training data set. The method also generates reliability estimates which makes it possible to distinguish between membrane-spanning proteins (I, intrinsic) and globular proteins with regions of high hydrophobicity buried in the core.

- 10 For a protein of length L , the block of length l with maximum hydrophobicity is found:

$$\max H = \max_{k=1, \dots, L-l+1} (1/l) \sum_{i=k}^{k+l-1} H_i$$

where H_i represents the hydrophobicity of an individual residue.

- 15 Let $P(I/\max H)$ and $P(E/\max H)$ be the conditional probabilities that a protein is integral or peripheral, respectively, given its value of maximal hydrophobicity $\max H$, and let $P(I)$ and $P(E)$ be the prior probabilities of intrinsic and extrinsic membrane proteins estimated from the training set. Then a sequence is assigned to E if

$$P(E/\max H) > P(I/\max H)$$

or, after applying the Bayes rule,

$$P(E)P(\max H/E) > P(I)P(\max H/I),$$

- 25 where the conditional probabilities $P(\max H/E)$ and $P(\max H/I)$ can be determined based on the estimates of probability distributions of $\max H$ in both groups.

- Discriminant analysis allows to simplify this task by calculating the odds $P(E/\max H):P(I/\max H)$ as e^b , where b is the left-hand side of a linear or quadratic inequality. For example, for the window of length 17, the protein is allocated to the

peripheral category E based on the empirically derived quadratic inequality:

$$1.05(\text{maxH})^2 + 12.30\text{maxH} + 17.49 > 0,$$

whereas the optimal inequality for assigning membrane
5 proteins (category I) is linear:

$$-9.02\text{maxH} + 14.27 > 0$$

The odds parameter can be made more or less stringent. For example, one can require odds at least 1:10 for a protein to be classified as integral. This leads to higher selectivity but less
10 sensitivity.

The boundaries of membrane-spanning regions in putative membrane proteins are detected by means of an iterative procedure whereby the most hydrophobic region corresponding to the value maxH is considered to be membrane and removed from the sequence.
15 The classification procedure is then repeated again for the remaining sequence, and, if such a protein is again classified as integral, the next most hydrophobic region is considered.

Reference: Klein, P., Kanehisa, M., DeLisi, C. (1985) The detection and classification of membrane-spanning proteins.
20 *Biochem Biophys Acta* 815: 468-476

Transcription factors

Purified eukaryotic RNA polymerase II is unable to initiate promoter-specific transcription. A family of factors that collectively confer RNAPII promoter specificity is known as the
25 general transcription factors (GTFs). They include the TATA-binding Protein (TBP) TFIIB, TFIIE, TFIIF and TFI IH. These factors are conserved among all eukaryotes.

RNAPII complexes containing the entire set of GTFs or a subset of GTFs together with other proteins have been isolated
30 from mammalian and yeast cells. Although purified RNAPII and GTFs are sufficient for promoter-specific initiation, this system fails to respond to activators. This is mediated by a further complex termed mediator complex which associates with the

carboxy-terminal heptapeptide domain (CTD) of the largest subunit of RNAPII.

Purification of human RNAPII complexes resulted in two distinct forms of human RNAPII after analysis of functional properties. One complex contained chromatin remodeling activities but was devoid of GTFs. The other complex did not contain factors that modify chromatin but contained a subset of SRB/mediator subunits and GTFs and other polypeptides that mediate transcriptional activation, a scenario similar to that reported for yeast.

A complex designated NAT (~20 SU) for negative regulator of transcription contains RNAPII, Cdk8, homologs of the yeast mediator complex as well as Rgr1 and Srb10/11 known as negative regulators of transcription.

A complex with striking similar structural and functional properties to NAT has been identified designated SMCC (~15 SU) (SRB/mediator coactivator complex), that can also mediate transcriptional activation.

The SMCC complex includes all reported NAT subunits including subunits of the TRAP complex. TRAP is a coactivator complex isolated on the basis of its interaction with the thyroid hormone receptor. Another coactivator complex DRIP, isolated on the basis of its ability to interact with the vitamin D3 receptor, contains novel subunits as well as subunits of NAT/SMCC and TRAP complexes.

The effects of each of these coactivator complexes is dependent on the TFIID complex. It is not known if the TAF subunits of TFIID are required. It is likely that new coactivator complexes will be uncovered containing both novel and previously defined components.

Beside the huge amount of transcription factors which can be part of the RNAIIP holoenzyme or the coactivator complexes there is an even larger quantity of specific transcription factors binding to promoter elements within the DNA sequences of a given gene leading to activation or repression of transcription. A

broad range of cellular responses like differentiation, proliferation, cell death and others are elicited through activating or repressing the transcription of target genes.

5 There are at least five superclasses of transcription factors:

1. Superclass contains members with characteristic basic domains:

Members are:

10 Leucine zipper factors, where the basic domain is followed by a leucine zipper of repeated leucine residues at every seventh position. The zipper mediates protein dimerization as a prerequisite for DNA-binding.

15 Helix-loop-helix factors (bHLH) contain a DNA-binding basic region followed by a motif of two potential amphipathic alpha-helices connected by a loop of variable length also mediating dimerization.

Factors with a combination of Helix-loop-helix and leucine zipper.

20 Further members of this superclass are NF-1, RF-X, and bHSH like proteins.

2. Superclass comprises factors containing zinc-coordinating DNA-binding domains.

Members are:

25 Proteins with Cys4 zinc finger of nuclear receptor type, where two such motifs differing in size, composition and function are present in each receptor molecule. Each finger comprises 4 cysteine residues coordinating one zinc ion. The second half including the second cysteine pair has alpha-helix conformation and the helix of the first finger binds to the DNA through the
30 major groove. The sequence between the first two cysteines of the second finger mediates dimerization upon DNA-binding. This class includes the steroid hormone receptors and the thyroid hormone

receptor-like factors. Other diverse cys4 zinc fingers have a motif of GATA-type.

Proteins with Cys2His2 zinc finger domain(s). Each finger comprises 2 cysteine and 2 histidine residues coordinating one zinc ion, and in some cases one histidine is replaced by another cysteine. The zinc ion is essential for DNA-binding.

Proteins with Cys6 cysteine-zinc cluster(s). Six cysteine residues coordinate two zinc ions, i. e. two of the thiol groups are coordinating two zinc ions each. Present in many fungal regulators.

Zinc fingers of alternating composition.

3. Superclass contains factors of helix-turn-helix type.

Members are:

Proteins with homeo domains. Homeo domains are three consecutive alpha-helix structures. Helix 3 contacts mainly the major groove of the DNA, some contacts at the minor groove are observed as well. Helix 2 and 3 resemble the helix-turn-helix structure of prokaryotic regulators.

Proteins with Paired box domain(s). This is a DNA-binding domain of approximately 130 amino acid residues. Its N-terminal half is basic, its C-terminal half is highly charged in general. It probably comprises 3 alpha-helices.

Proteins with Fork head / winged helix domain(s). This domain was identified by homology between HNF-3A and fkh. The domain comprises approx. 110 AA. Analysis of the crystal structure has revealed a compact structure of three alpha-helices, the third alpha-helix being exposed towards the major groove of the DNA. The domain also exerts minor groove contacts. Upon binding to DNA, it induces a bend of 13 degree.

Heat shock factors

Proteins with Tryptophan clusters. The tryptophan clusters comprise several tryptophan residues with a spacing of 12-21

amino acid residues; the subclass of myb-type DNA-binding domains typically exhibit a spacing of 19-21 amino acid residues.

Proteins with TEA domain(s). The TEA domain has been identified as a region which is conserved among the transcription factors TEF-1, TECl and abaA. This domain in TEF-1 has been shown to interact with DNA, although two additional regions may also contribute to DNA-binding. It is predicted to fold into three alpha-helices, with a randomly coiled region of 16-18 amino acid residues between helices 1 and 2, and a short stretch between helices 2 and 3 of 3-8 residues.

4. Superclass contains beta-Scaffold Factors with Minor Groove Contacts

Members are:

Proteins with RHR (Rel homology) region.

The structure of the Rel-type DBD exhibits a bipartite subdomain structure, each subdomain comprising a beta-barrel with five loops that form an extensive contact surface to the major groove of the DNA. Particularly, the first loop of the N-terminal subdomain (the highly conserved recognition loop) performs contacts with the recognition element on the DNA, but other loops are involved. The fact that the main DNA-contacts are made through loops has been suggested to provide a high degree of flexibility in binding to a range of different target sequences. Augmenting interactions are achieved by two alpha-helices within the N-terminal Part that form strong minor groove contacts to the A/T-rich center of the B-element. In pb5, the sequence between both alpha-helices is much shorter and even helix 2 is truncated. The second, C-terminal domain is necessary mainly for protein dimerization.

p53 proteins

MADS (MCM1-agamous-deficiens-SRF) box proteins. Proteins of this class comprise a region of homology. The DNA-binding domain also comprises the dimerization capability. In the DNA-bound dimer (shown for SRF), two antiparallel amphipathic alpha-helices

(alpha-I), form a coiled coil and are oriented approximately parallel on the minor groove. These helices make minor and major groove contacts, the N-terminal extensions form minor groove contacts. The bound DNA is bent and wrapped around the protein. 5 It exhibits a compressed minor groove in the center and widened minor groove in the flanks.

Beta-Barrel alpha-helix transcription factors.

TATA-binding proteins

HMG proteins

10 Proteins of this class comprise a region of homology with the chromosomal non-histone HMG proteins such as HMGL. This region comprises the DNA-binding domain which in some instances such as HMGL mediates sequence-unspecific, in other cases such LEF-1 sequence-specific binding to DNA. This domain exhibits a 15 typical L-shaped conformation made up of 3 alpha-helices and an extended N-terminal extension of the first helix. The latter together with helix 1, which contains a kink, form the long arm of the L, whereas helices 1 and 2 form the short arm. Binding to the minor groove induces a sharp bending of the DNA by more than 20 90 degree, away from the bound protein. The overall topology of the DNA-protein complexes resembles somewhat that of the TBP-TATA box complex.

Heteromeric CCAAT factors

Proteins with Grainyhead domain(s)

25 Cold-shock domain factors. Cold-shock domain proteins are characterized by a highly conserved region first found in prokaryotic cold-shock proteins. This domain is a single-stranded nucleic acid-binding structure interacting with DNA or RNA. It consists of an antiparallel five-stranded beta-barrel, the 30 strands of which are connected by turns and loops. Within this structure, a three-stranded beta-strand contains a conserved RNA-binding motif, RNP1. Not all CSD proteins are transcription factors. Those which specifically bind to a certain sequence are termed Y-box proteins. Proteins of this class were previously

called protamine-like domain proteins because of having a highly positively charged domain with interspersed proline residues.

Proteins with Runt homology domain

5 The members of this transcription factor class have been identified on the basis of their homology to a defined region within the *Drosophila* protein Runt. The runt domain is part of the DNA-binding domain of these factors. It consists mainly of beta-strands, does not contain alpha-helical regions and seems to be most similar to the palm domain found in DNA polymerase beta
10 (rat).

5. Superclass contains other transcription factors like Copper fist proteins, HMGI(Y), STAT, Pocket domain proteins and Ap2/EREBP-related factors.

15 The classification of transcription factors originates from TRANSFAC database:

<http://transfac.gbf.de/TRANSFAC/>

Reference: Heinemeyer

Several categories of proteins are coded for by clones of the invention within the overall group of "Transcription Factors"
20 and include, among others, the following:

Homeobox-proteins: Homeodomain-containing transcription factors are essential for a variety of processes in vertebrate development, including organogenesis. They have been shown to
25 regulate cell proliferation, pattern segmental identity and determine cell fate decisions during embryogenesis. For example, In zebrafish *emx2* mRNAs are found in the dorsal telencephalon, parts of the diencephalon and the otocyst. The human homologue *Emx2* appears to be already expressed in 8.5 day
30 embryos. It is also expressed in the presumptive cerebral cortex, olfactory bulbs, in some neuroectodermal areas in embryonic head including olfactory placodes in earlier stages and olfactory epithelia later in development. Mutants of the *D. melanogaster* gene "empty spiracles" display spiracles devoid of filzkörper.

no antenna and an open head. Clones in this category include:
amy2_14m1b.

Proteins with myc-type, helix-loop-helix dimerization domain signature(s). This helix-loop-helix domain mediates protein
5 dimerization has been found in various multimeric transcription factors. Clones in this category include: tes3_18n14.

Transcriptional silencers: In addition to transcription factors, other proteins, such as YDL153c of *Saccharomyces cerevisia* are responsible for silencing of genes. Clones in this
10 category include: amy2_2f22.

Proteins regulating transcription factors: The activity of several transcription factor is regulated by the binding or dissociation of other proteins or by phosphorylation or
dephosphorylation of the transcription factor. For example, I-
15 kappa-B-related protein interacts with the transcription factor NF-kB. I-kappa-B-alpha mutations contribute to constitutive NF-kappaB activity in cultured and primary HRS (Hodgkin/Reed-Sternberg) cells and are therefore involved in the pathogenesis of Hodgkin's disease (HD) patients. Clones in this category
20 include: amy2_1c12.

Signal transducing proteins: Beta-transducin subunits of G-proteins contain WD-40 repeats. The beta subunits seem to be required for the replacement of GDP by GTP as well as for
membrane anchoring and receptor recognition. Due to the zinc
25 finger the novel protein seems to be a new molecule involved in signal transduction and transcription. These proteins have been reported by OMIM to be associated (as potentially diagnostic, therapeutic, causative, and/or related, etc...) with the following diseases: 1) essential hypertension (OMIM *139130). Clones in
30 this category include: tes3_11c22.

* * *

The invention, therefore, specifically contemplates the following assemblages of materials, which track the above-identified fourteen functional groupings, that are useful in
35 practicing the profiling aspects of the invention. One type of assemblage is nucleic acid-based and can include the following groupings of sequences and their derivatives: all sequences; human fetal brain sequences; brain derived sequences; human fetal

kidney library sequences; kidney derived sequences; human mammary carcinoma library sequences; mammary carcinoma derived sequences; human testis library sequences; testes derived sequences; cell cycle genes; cell structure and motility genes; differentiation and development genes; intracellular transport and trafficking genes; metabolism genes; nucleic acid management genes; signal transduction genes; transmembrane protein genes; and transcription factor genes. Other assemblages contain proteins or their corresponding antibodies or antibody fragments, divided along the same groupings.

Database Applications

Because they are human genes and gene products, the inventive molecules are useful as members of a database. Such a database may be used, for example, in drug discovery and rationale drug design or in testing the novelty and non-obviousness of newly sequenced materials. In addition, they are particularly suited in designing variants for the profiling (and other) applications described herein. Hence, the following discussion of electronic embodiments applies equally to such variants, which, naturally, will be generated and stored using a computer using known methodologies.

Accordingly, one aspect of the invention contemplates a database of at least one of the inventive sequences stored on computer readable media. Again, the individual sequences may be grouped with regard to the individual functional and structural groups mentioned above. While the individual sequences of a database may exist in printed form, they are preferably in electronic form, as in an ascii or a text file. They may also exist as word processing files or they may be stored in database applications like DB2, Sybase, Oracle, GCG and GenBank. One skilled in the art will understand the range of applications suitable for using and storing the electronic embodiments of the invention.

"Computer readable media" refers to any medium which can be read and accessed by a computer. These include: magnetic storage media, like floppy discs, hard drives and magnetic tape; optical storage media, like CD-ROM; electrical storage media, like RAM and ROM; and hybrids of these categories, like magnetic/optical

storage media. One skilled in the art will readily understand the scope of computer readable media and how to implement them.

Biological Activities and Assays for Implementing Therapeutic and Diagnostic Applications

5 This section provides assays for biological activity that are useful in characterizing and quantifying the biological activity of the inventive molecules and their derivatives, which is relevant to the pharmacological effects of the inventive molecules. As used in this section, it will be understood that
10 "protein" may also refer to the inventive antibodies (including fragments).

Cytokine and Cell Proliferation/Differentiation Activity

A protein of the present invention may exhibit cytokine, cell proliferation (either inducing or inhibiting) or cell
15 differentiation (either inducing or inhibiting) activity or may induce production of other cytokines in certain cell populations. Many protein factors discovered to date, including all known cytokines, have exhibited activity in one or more factor dependent cell proliferation assays, and hence the assays serve
20 as a convenient confirmation of cytokine activity. The activity of a protein of the present invention is evidenced by any one of a number of routine factor dependent cell proliferation assays for cell lines including, without limitation, 32D, DA2, DA16, T10, B9, B9/11, BaF3, MC9/G, M + (preB M +), 2E8, RB5, DA1, 123,
25 T11b5, HT2, CTLL2, TF-1, Mo7e and CMK.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for T-cell or thymocyte proliferation include without limitation those described in: Current Protocols in Immunology,
30 Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, In Vitro assays for Mouse Lymphocyte Function 3.1-3.19; Chapter 7, Immunologic studies in Humans); Takai et al., J. Immunol. 137:3494-3500, 1986; Bertagnolli et
35 al., J. Immunol. 145:1706-1712, 1990; Bertagnolli et al., Cellular Immunology 133:327-341, 1991; Bertagnolli, et al., I. Immunol. 149:3778-3783, 1992; Bowman et al., I. Immunol. 152:1756-1761, 1994.

Assays for cytokine production and/or proliferation of spleen cells, lymph node cells or thymocytes include, without limitation, those described in: Polyclonal T cell stimulation, Kruisbeek, A. M. and Shevach, E. M. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 3.12.1-3.12.14, John Wiley and Sons, Toronto. 1994; and Measurement of mouse and human interleukin gamma, Schreiber, R. D. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 6.8.1-6.8.8, John Wiley and Sons, Toronto. 1994.

Assays for proliferation and differentiation of hematopoietic and lymphopoietic cells include, without limitation, those described in: Measurement of Human and Murine Interleukin 2 and Interleukin 4, Bottomly, K., Davis, L. S. and Lipsky, P. E. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 6.3.1-6.3.12, John Wiley and Sons, Toronto. 1991; deVries et al., J. Exp. Med. 173:1205-1211, 1991; Moreau et al., Nature 336:690-692, 1988; Greenberger et al., Proc. Natl. Acad. Sci. U.S.A. 80:2931-2938, 1983; Measurement of mouse and human interleukin 6-Nordan, R. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 6.6.1-6.6.5, John Wiley and Sons, Toronto. 1991; Smith et al., Proc. Natl. Acad. Sci. U.S.A. 83:1857-1861, 1986; Measurement of human Interleukin 11-Bennett, F., Giannotti, J., Clark, S. C. and Turner, K. J. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 6.15.1 John Wiley and Sons, Toronto. 1991; Measurement of mouse and human Interleukin 9-Ciarletta, A., Giannotti, J., Clark, S. C. and Turner, K. J. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 6.13.1, John Wiley and Sons, Toronto. 1991.

Assays for T-cell clone responses to antigens (which will identify, among others, proteins that affect APC-T cell interactions as well as direct T-cell effects by measuring proliferation and cytokine production) include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, In Vitro assays for Mouse Lymphocyte Function; Chapter 6, Cytokines and their cellular receptors; Chapter 7, Immunologic studies in Humans); Weinberger et al.,

Proc. Natl. Acad. Sci. USA 77:6091-6095, 1980; Weinberger et al., Eur. J. Immun. 11:405-411, 1981; Takai et al., J. Immunol. 137:3494-3500, 1986; Takai et al., J. Immunol. 140:508-512, 1988.

Immune Stimulating or Suppressing Activity

5 A protein of the present invention may also exhibit immune stimulating or immune suppressing activity, including without limitation the activities for which assays are described herein. A protein may be useful in the treatment of various immune deficiencies and disorders (including severe combined
10 immunodeficiency (SCID)), e.g., in regulating (up or down) growth and proliferation of T and/or B lymphocytes, as well as effecting the cytolytic activity of NK cells and other cell populations. These immune deficiencies may be genetic or be caused by vital (e.g., HIV) as well as bacterial or fungal infections, or may
15 result from autoimmune disorders. More specifically, infectious diseases caused by viral, bacterial, fungal or other infection may be treatable using a protein of the present invention, including infections by HIV, hepatitis viruses, herpesviruses, mycobacteria, Leishmania spp., malaria spp. and various fungal
20 infections such as candidiasis. Of course, in this regard, a protein of the present invention may also be useful where a boost to the immune system generally may be desirable, i.e., in the treatment of cancer.

Autoimmune disorders which may be treated using a protein of
25 the present invention include, for example, connective tissue disease, multiple sclerosis, systemic lupus erythematosus, rheumatoid arthritis, autoimmune pulmonary inflammation, Guillain-Barre syndrome, autoimmune thyroiditis, insulin dependent diabetes mellitus, myasthenia gravis, graft-versus-host
30 disease and autoimmune inflammatory eye disease. Such a protein of the present invention may also be useful in the treatment of allergic reactions and conditions, such as asthma (particularly allergic asthma) or other respiratory problems. Other conditions, in which immune suppression is desired
35 (including, for example, organ transplantation), may also be treatable using a protein of the present invention.

Using the proteins of the invention it may also be possible to modify immune responses, in a number of ways. Down regulation

may be in the form of inhibiting or blocking an immune response already in progress or may involve preventing the induction of an immune response. The functions of activated T cells may be inhibited by suppressing T cell responses or by inducing specific tolerance in T cells, or both. Immunosuppression of T cell responses is generally an active, non-antigen-specific, process which requires continuous exposure of the T cells to the suppressive agent. Tolerance, which involves inducing non-responsiveness or anergy in T cells, is distinguishable from immunosuppression in that it is generally antigen-specific and persists after exposure to the tolerizing agent has ceased. Operationally, tolerance can be demonstrated by the lack of a T cell response upon reexposure to specific antigen in the absence of the tolerizing agent.

Down regulating or preventing one or more antigen functions (including without limitation B lymphocyte antigen functions (such as, for example, B7)), e.g., preventing high level lymphokine synthesis by activated T cells, will be useful in situations of tissue, skin and organ transplantation and in graft-versus-host disease (GVHD). For example, blockage of T cell function should result in reduced tissue destruction in tissue transplantation. Typically, in tissue transplants, rejection of the transplant is initiated through its recognition as foreign by T cells, followed by an immune reaction that destroys the transplant. The administration of a molecule which inhibits or blocks interaction of a B7 lymphocyte antigen with its natural ligand(s) on immune cells (such as a soluble, monomeric form of a peptide having B7-2 activity alone or in conjunction with a monomeric form of a peptide having an activity of another B lymphocyte antigen (e.g., B7-1, B7-3) or blocking antibody), prior to transplantation can lead to the binding of the molecule to the natural ligand(s) on the immune cells without transmitting the corresponding costimulatory signal. Blocking B lymphocyte antigen function in this manner prevents cytokine synthesis by immune cells, such as T cells, and thus acts as an immunosuppressant. Moreover, the lack of costimulation may also be sufficient to anergize the T cells, thereby inducing tolerance in a subject. Induction of long-term tolerance by B lymphocyte antigen-blocking reagents may avoid the necessity of repeated

administration of these blocking reagents. To achieve sufficient immunosuppression or tolerance in a subject, it may also be necessary to block the function of a combination of B lymphocyte antigens.

- 5 The efficacy of particular blocking reagents in preventing organ transplant rejection or GVHD can be assessed using animal models that are predictive of efficacy in humans. Examples of appropriate systems which can be used include allogeneic cardiac grafts in rats and xenogeneic pancreatic islet cell grafts in
10 mice, both of which have been used to examine the immunosuppressive effects of CTLA4Ig fusion proteins in vivo as described in Lenschow et al., Science 257:789-792 (1992) and Turka et al., Proc. Natl. Acad. Sci USA, 89:11102-11105 (1992). In addition, murine models of GVHD (see Paul ed., Fundamental
15 Immunology, Raven Press, New York, 1989, pp. 846-847) can be used to determine the effect of blocking B lymphocyte antigen function in vivo on the development of that disease.

- Blocking antigen function may also be therapeutically useful for treating autoimmune diseases. Many autoimmune disorders are
20 the result of inappropriate activation of T cells that are reactive against self tissue and which promote the production of cytokines and autoantibodies involved in the pathology of the diseases. Preventing the activation of autoreactive T cells may reduce or eliminate disease symptoms. Administration of reagents
25 which block costimulation of T cells by disrupting receptor:ligand interactions of B lymphocyte antigens can be used to inhibit T cell activation and prevent production of autoantibodies or T cell-derived cytokines which may be involved in the disease process. Additionally, blocking reagents may
30 induce antigen-specific tolerance of autoreactive T cells which could lead to long-term relief from the disease. The efficacy of blocking reagents in preventing or alleviating autoimmune disorders can be determined using a number of well-characterized animal models of human autoimmune diseases. Examples include
35 murine experimental autoimmune encephalitis, systemic lupus erythmatosis in MRL/lpr/lpr mice or NZB hybrid mice, murine autoimmune collagen arthritis, diabetes mellitus in NOD mice and BB rats, and murine experimental myasthenia gravis (see Paul ed.,

Fundamental Immunology, Raven Press, New York, 1989, pp. 840-856).

Upregulation of an antigen function (preferably a B lymphocyte antigen function), as a means of up regulating immune responses, may also be useful in therapy. Upregulation of immune responses may be in the form of enhancing an existing immune response or eliciting an initial immune response. For example, enhancing an immune response through stimulating B lymphocyte antigen function may be useful in cases of viral infection. In addition, systemic viral diseases such as influenza, the common cold, and encephalitis might be alleviated by the administration of stimulatory forms of B lymphocyte antigens systemically.

Alternatively, anti-viral immune responses may be enhanced in an infected patient by removing T cells from the patient, costimulating the T cells in vitro with viral antigen-pulsed APCs either expressing a peptide of the present invention or together with a stimulatory form of a soluble peptide of the present invention and reintroducing the in vitro activated T cells into the patient. Another method of enhancing anti-viral immune responses would be to isolate infected cells from a patient, transfect them with a nucleic acid encoding a protein of the present invention as described herein such that the cells express all or a portion of the protein on their surface, and reintroduce the transfected cells into the patient. The infected cells would now be capable of delivering a costimulatory signal to, and thereby activate, T cells in vivo.

In another application, up regulation or enhancement of antigen function (preferably B lymphocyte antigen function) may be useful in the induction of tumor immunity. Tumor cells (e.g., sarcoma, melanoma, lymphoma, leukemia, neuroblastoma, carcinoma) transfected with a nucleic acid encoding at least one peptide of the present invention can be administered to a subject to overcome tumor-specific tolerance in the subject. If desired, the tumor cell can be transfected to express a combination of peptides. For example, tumor cells obtained from a patient can be transfected ex vivo with an expression vector directing the expression of a peptide having B7-2-like activity alone, or in conjunction with a peptide having B7-1-like activity and/or B7-3-like activity. The transfected tumor cells are returned to the

patient to result in expression of the peptides on the surface of the transfected cell. Alternatively, gene therapy techniques can be used to target a tumor cell for transfection in vivo.

The presence of the peptide of the present invention having the activity of a B lymphocyte antigen(s) on the surface of the tumor cell provides the necessary costimulation signal to T cells to induce a T cell mediated immune response against the transfected tumor cells. In addition, tumor cells which lack MHC class I or MHC class II molecules, or which fail to reexpress sufficient mounts of MHC class I or MHC class II molecules, can be transfected with nucleic acid encoding all or a portion of (e.g., a cytoplasmic-domain truncated portion) of an MHC class I alpha chain protein and beta 2 microglobulin protein or an MHC class II alpha chain protein and an MHC class II beta chain protein to thereby express MHC class I or MHC class II proteins on the cell surface. Expression of the appropriate class I or class II MHC in conjunction with a peptide having the activity of a B lymphocyte antigen (e.g., B7-1, B7-2, B7-3) induces a T cell mediated immune response against the transfected tumor cell.

Optionally, a gene encoding an antisense construct which blocks expression of an MHC class II associated protein, such as the invariant chain, can also be cotransfected with a DNA encoding a peptide having the activity of a B lymphocyte antigen to promote presentation of tumor associated antigens and induce tumor specific immunity. Thus, the induction of a T cell mediated immune response in a human subject may be sufficient to overcome tumor-specific tolerance in the subject.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Suitable assays for thymocyte or splenocyte cytotoxicity include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, In Vitro assays for Mouse Lymphocyte Function 3.1-3.19; Chapter 7, Immunologic studies in Humans); Herrmann et al., Proc. Natl. Acad. Sci. USA 78:2488-2492, 1981; Herrmann et al., J. Immunol. 128:1968-1974, 1982; Handa et al., J. Immunol. 135:1564-1572, 1985; Takai et al., I. Immunol. 137:3494-3500, 1986; Takai et al., J. Immunol.

140:508-512, 1988; Herrmann et al., Proc. Natl. Acad. Sci. USA
78:2488-2492, 1981; Herrmann et al., J. Immunol. 128:1968-1974,
1982; Handa et al., J. Immunol. 135:1564-1572, 1985; Takai et
al., J. Immunol. 137:3494-3500, 1986; Bowman et al., J. Virology
5 61:1992-1998; Takai et al., J. Immunol. 140:508-512, 1988;
Bertagnolli et al., Cellular Immunology 133:327-341, 1991; Brown
et al., J. Immunol. 153:3079-3092, 1994.

Assays for T-cell-dependent immunoglobulin responses and
isotype switching (which will identify, among others, proteins
10 that modulate T-cell dependent antibody responses and that affect
Th1/Th2 profiles) include, without limitation, those described
in: Maliszewski, J. Immunol. 144:3028-3033, 1990; and Assays for
B cell function: In vitro antibody production, Mond, J. J. and
Brunswick, M. In Current Protocols in Immunology. J. E. e.a.
15 Coligan eds. Vol 1-pp. 3.8.1-3.8.1b, John Wiley and Sons,
Toronto. 1994.

Mixed lymphocyte reaction (MLR) assays (which will identify,
among others, proteins that generate predominantly Th1 and CTL
responses) include, without limitation, those described in:
20 Current Protocols in Immunology, Ed by J. E. Coligan, A. M.
Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub.
Greene Publishing Associates and Wiley-Interscience (Chapter 3,
In Vitro assays for Mouse Lymphocyte Function 3.1-3.19; Chapter
7, Immunologic studies in Humans); Takai et al., J. Immunol.
25 137:3494-3500, 1986; Takai et al., J. Immunol. 140:508-512, 1988;
Bertagnolli et al., J. Immunol. 149:3778-3783, 1992.

Dendritic cell-dependent assays (which will identify, among
others, proteins expressed by dendritic cells that activate naive
T-cells) include, without limitation, those described in: Guery
30 et al., J. Immunol. 134:536-544, 1995; Inaba et al., Journal of
Experimental Medicine 173:549-559, 1991; Macatonia et al.,
Journal of Immunology 154:5071-5079, 1995; Porgador et al.,
Journal of Experimental Medicine 182:255-260, 1995; Nair et al.,
Journal of Virology 67:4062-4069, 1993; Huang et al., Science
35 264:961-965, 1994; Macatonia et al., Journal of Experimental
Medicine 169:1255-1264, 1989; Bhardwaj et al., Journal of
Clinical Investigation 94:797-807, 1994; and Inaba et al.,
Journal of Experimental Medicine 172:631-640, 1990.

Assays for lymphocyte survival/apoptosis (which will identify, among others, proteins that prevent apoptosis after superantigen induction and proteins that regulate lymphocyte homeostasis) include, without limitation, those described in:

- 5 Darzynkiewicz et al., Cytometry 13:795-808, 1992; Gorczyca et al., Leukemia 7:659-670, 1993; Gorczyca et al., Cancer Research 53:1945-1951, 1993; Itoh et al., Cell 66:233-243, 1991; Zacharchuk, Journal of Immunology 145:4037-4045, 1990; Zamai et al., Cytometry 14:891-897, 1993; Gorczyca et al., International
10 Journal of Oncology 1:639-648, 1992.

Assays for proteins that influence early steps of T-cell commitment and development include, without limitation, those described in: Antica et al., Blood 84:111-117, 1994; Fine et al., Cellular Immunology 155:111-122, 1994; Galy et al., Blood
15 85:2770-2778, 1995; Toki et al., Proc. Nat. Acad Sci. USA 88:7548-7551, 1991.

Hematopoiesis Regulating Activity

A protein of the present invention may be useful in regulation of hematopoiesis and, consequently, in the treatment of myeloid or lymphoid cell deficiencies. Even marginal
5 biological activity in support of colony forming cells or of factor-dependent cell lines indicates involvement in regulating hematopoiesis, e.g. in supporting the growth and proliferation of erythroid progenitor cells alone or in combination with other cytokines, thereby indicating utility, for example, in treating
10 various anemias or for use in conjunction with irradiation/chemotherapy to stimulate the production of erythroid precursors and/or erythroid cells; in supporting the growth and proliferation of myeloid cells such as granulocytes and monocytes/macrophages (i.e., traditional CSF activity) useful,
15 for example, in conjunction with chemotherapy to prevent or treat consequent myelo-suppression; in supporting the growth and proliferation of megakaryocytes and consequently of platelets thereby allowing prevention or treatment of various platelet disorders such as thrombocytopenia, and generally for use in
20 place of or complimentary to platelet transfusions; and/or in supporting the growth and proliferation of hematopoietic stem cells which are capable of maturing to any and all of the above-mentioned hematopoietic cells and therefore find therapeutic utility in various stem cell disorders (such as those usually
25 treated with transplantation, including, without limitation, aplastic anemia and paroxysmal nocturnal hemoglobinuria), as well as in repopulating the stem cell compartment post irradiation/chemotherapy, either in-vivo or ex-vivo (i.e., in conjunction with bone marrow transplantation or with peripheral
30 progenitor cell transplantation (homologous or heterologous)) as normal cells or genetically manipulated for gene therapy.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Suitable assays for proliferation and differentiation of
35 various hematopoietic lines are cited above.

Assays for embryonic stem cell differentiation (which will identify, among others, proteins that influence embryonic differentiation hematopoiesis) include, without limitation, those described in: Johansson et al. Cellular Biology 15:141-151, 1995;

Keller et al., Molecular and Cellular Biology 13:473-486, 1993;
McClanahan et al., Blood 81:2903-2915, 1993.

Assays for stem cell survival and differentiation (which will identify, among others, proteins that regulate lympho-
5 hematopoiesis) include, without limitation, those described in:
Methylcellulose colony forming assays, Freshney, M. G. In Culture
of Hematopoietic Cells. R. I. Freshney, et al. eds. Vol pp. 265-
268, Wiley-Liss, Inc., New York, N.Y. 1994; Hirayama et al.,
Proc. Natl. Acad. Sci. USA 89:5907-5911, 1992; Primitive
10 hematopoietic colony forming cells with high proliferative
potential, McNiece, I. K. and Briddell, R. A. In Culture of
Hematopoietic Cells. R. I. Freshney, et al. eds. Vol pp. 23-39,
Wiley-Liss, Inc., New York, N.Y. 1994; Neben et al., Experimental
Hematology 22:353-359, 1994; Cobblestone area forming cell assay,
15 Ploemacher, R. E. In Culture of Hematopoietic Cells. R. I.
Freshney, et al. eds. Vol pp. 1-21, Wiley-Liss, Inc., New York,
N.Y. 1994; Long term bone marrow cultures in the presence of
stromal cells, Spooncer, E., Dexter, M. and Allen, T. In Culture
of Hematopoietic Cells. R. I. Freshney, et al. eds. Vol pp. 163-
20 179, Wiley-Liss, Inc., New York, N.Y. 1994; Long term culture
initiating cell assay, Sutherland, H. J. In Culture of
Hematopoietic Cells. R. I. Freshney, et al. eds. Vol pp. 139-162,
Wiley-Liss, Inc., New York, N.Y. 1994.

Tissue Growth Activity

25 A protein of the present invention also may have utility in
compositions used for bone, cartilage, tendon, ligament and/or
nerve tissue growth or regeneration, as well as for wound healing
and tissue repair and replacement, and in the treatment of burns,
incisions and ulcers.

30 A protein of the present invention, which induces cartilage
and/or bone growth in circumstances where bone is not normally
formed, has application in the healing of bone fractures and
cartilage damage or defects in humans and other animals. Such a
preparation employing a protein of the invention may have
35 prophylactic use in closed as well as open fracture reduction and
also in the improved fixation of artificial joints. De novo bone
formation induced by an osteogenic agent contributes to the
repair of congenital, trauma induced, or oncologic resection

induced craniofacial defects, and also is useful in cosmetic plastic surgery.

A protein of this invention may also be used in the treatment of periodontal disease, and in other tooth repair processes. Such agents may provide an environment to attract bone-forming cells, stimulate growth of bone-forming cells or induce differentiation of progenitors of bone-forming cells. A protein of the invention may also be useful in the treatment of osteoporosis or osteoarthritis, such as through stimulation of bone and/or cartilage repair or by blocking inflammation or processes of tissue destruction (collagenase activity, osteoclast activity, etc.) mediated by inflammatory processes.

Another category of tissue regeneration activity that may be attributable to the protein of the present invention is tendon/ligament formation. A protein of the present invention, which induces tendon/ligament-like tissue or other tissue formation in circumstances where such tissue is not normally formed, has application in the healing of tendon or ligament tears, deformities and other tendon or ligament defects in humans and other animals. Such a preparation employing a tendon/ligament-like tissue inducing protein may have prophylactic use in preventing damage to tendon or ligament tissue, as well as use in the improved fixation of tendon or ligament to bone or other tissues, and in repairing defects to tendon or ligament tissue. De novo tendon/ligament-like tissue formation induced by a composition of the present invention contributes to the repair of congenital, trauma induced, or other tendon or ligament defects of other origin, and is also useful in cosmetic plastic surgery for attachment or repair of tendons or ligaments. The compositions of the present invention may provide environment to attract tendon- or ligament-forming cells, stimulate growth of tendon- or ligament-forming cells, induce differentiation of progenitors of tendon- or ligament-forming cells, or induce growth of tendon/ligament cells or progenitors ex vivo for return in vivo to effect tissue repair. The compositions of the invention may also be useful in the treatment of tendonitis, carpal tunnel syndrome and other tendon or ligament defects. The compositions may also include an

appropriate matrix and/or sequestering agent as a carrier as is well known in the art.

The protein of the present invention may also be useful for proliferation of neural cells and for regeneration of nerve and brain tissue, i.e. for the treatment of central and peripheral nervous system diseases and neuropathies, as well as mechanical and traumatic disorders, which involve degeneration, death or trauma to neural cells or nerve tissue. More specifically, a protein may be used in the treatment of diseases of the peripheral nervous system, such as peripheral nerve injuries, peripheral neuropathy and localized neuropathies, and central nervous system diseases, such as Alzheimer's, Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis, and Shy-Drager syndrome. Further conditions which may be treated in accordance with the present invention include mechanical and traumatic disorders, such as spinal cord disorders, head trauma and cerebrovascular diseases such as stroke. Peripheral neuropathies resulting from chemotherapy or other medical therapies may also be treatable using a protein of the invention.

Proteins of the invention may also be useful to promote better or faster closure of non-healing wounds, including without limitation pressure ulcers, ulcers associated with vascular insufficiency, surgical and traumatic wounds, and the like.

It is expected that a protein of the present invention may also exhibit activity for generation or regeneration of other tissues, such as organs (including, for example, pancreas, liver, intestine, kidney, skin, endothelium), muscle (smooth, skeletal or cardiac) and vascular (including vascular endothelium) tissue, or for promoting the growth of cells comprising such tissues.

Part of the desired effects may be by inhibition or modulation of fibrotic scarring to allow normal tissue to regenerate. A protein of the invention may also exhibit angiogenic activity.

A protein of the present invention may also be useful for gut protection or regeneration and treatment of lung or liver fibrosis, reperfusion injury in various tissues, and conditions resulting from systemic cytokine damage.

A protein of the present invention may also be useful for promoting or inhibiting differentiation of tissues described

above from precursor tissues or cells; or for inhibiting the growth of tissues described above.

The activity of a protein of the invention may, among other means, be measured by the following methods:

5 Assays for tissue generation activity include, without limitation, those described in: International Patent Publication No. W095/16035 (bone, cartilage, tendon); International Patent Publication No. W095/05846 (nerve, neuronal); International Patent Publication No. W091/07491 (skin, endothelium).

10 Assays for wound healing activity include, without limitation, those described in: Winter, Epidermal Wound Healing, pps. 71-112 (Maibach, H. I. and Rovee, D. T., eds.), Year Book Medical Publishers, Inc., Chicago, as modified by Eaglstein and Mertz, J. Invest. Dermatol 71:382-84 (1978).

15 Activin/Inhibin Activity

 A protein of the present invention may also exhibit activin- or inhibin-related activities. Inhibins are characterized by their ability to inhibit the release of follicle stimulating hormone (FSH), while activins are characterized by their
20 ability to stimulate the release of follicle stimulating hormone (FSH). Thus, a protein of the present invention, alone or in heterodimers with a member of the inhibin alpha family, may be useful as a contraceptive based on the ability of inhibins to decrease fertility in female mammals and decrease spermatogenesis
25 in male mammals. Administration of sufficient amounts of other inhibins can induce infertility in these mammals. Alternatively, the protein of the invention, as a homodimer or as a heterodimer with other protein subunits of the inhibin- beta group, may be useful as a fertility inducing therapeutic, based upon the
30 ability of activin molecules in stimulating FSH release from cells of the anterior pituitary. See, for example, U.S. Pat. No. 4,798,885. A protein of the invention may also be useful for advancement of the onset of fertility in sexually immature mammals, so as to increase the lifetime reproductive performance
35 of domestic animals such as cows, sheep and pigs.

 The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for activin/inhibin activity include, without limitation, those described in: Vale et al., Endocrinology 91:562-572, 1972; Ling et al., Nature 321:779-782, 1986; Vale et al., Nature 321:776-779, 1986; Mason et al., Nature 318:659-663, 1985; Forage et al., Proc. Natl. Acad. Sci. USA 83:3091-3095, 1986.

Chemotactic/Chemokinetic Activity

A protein of the present invention may have chemotactic or chemokinetic activity (e.g., act as a chemokine) for mammalian cells, including, for example, monocytes, fibroblasts, neutrophils, T-cells, mast cells, eosinophils, epithelial and/or endothelial cells. Chemotactic and chemokinetic proteins can be used to mobilize or attract a desired cell population to a desired site of action. Chemotactic or chemokinetic proteins provide particular advantages in treatment of wounds and other trauma to tissues, as well as in treatment of localized infections. For example, attraction of lymphocytes, monocytes or neutrophils to tumors or sites of infection may result in improved immune responses against the tumor or infecting agent.

A protein or peptide has chemotactic activity for a particular cell population if it can stimulate, directly or indirectly, the directed orientation or movement of such cell population. Preferably, the protein or peptide has the ability to directly stimulate directed movement of cells. Whether a particular protein has chemotactic activity for a population of cells can be readily determined by employing such protein or peptide in any known assay for cell chemotaxis.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for chemotactic activity (which will identify proteins that induce or prevent chemotaxis) consist of assays that measure the ability of a protein to induce the migration of cells across a membrane as well as the ability of a protein to induce the adhesion of one cell population to another cell population. Suitable assays for movement and adhesion include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Marguiles, E. M. Shevach, W. Strober, Pub. Greene Publishing

Associates and Wiley-Interscience (Chapter 6.12, Measurement of
alpha and beta Chemokines 6.12.1-6.12.28; Taub et al. J. Clin.
Invest. 95:1370-1376, 1995; Lind et al. APMIS 103:140-146, 1995;
Muller et al Eur. J. Immunol. 25:1744-1748; Gruber et al. J. of
5 Immunol. 152:5860-5867, 1994; Johnston et al. J. of Immunol.
153:1762-1768, 1994.

Hemostatic and Thrombolytic Activity

A protein of the invention may also exhibit hemostatic or
thrombolytic activity. As a result, such a protein is expected to
10 be useful in treatment of various coagulation disorders
(including hereditary disorders, such as hemophilias) or to
enhance coagulation and other hemostatic events in treating
wounds resulting from trauma, surgery or other causes. A protein
of the invention may also be useful for dissolving or inhibiting
15 formation of thromboses and for treatment and prevention of
conditions resulting therefrom (such as, for example, infarction
of cardiac and central nervous system vessels (e.g., stroke).

The activity of a protein of the invention may, among other
means, be measured by the following methods:

20 Assay for hemostatic and thrombolytic activity include,
without limitation, those described in: Linet et al., J. Clin.
Pharmacol. 26:131-140, 1986; Burdick et al., Thrombosis Res.
45:413-419, 1987; Humphrey et al., Fibrinolysis 5:71-79 (1991);
Schaub, Prostaglandins 35:467-474, 1988.

25 Receptor/Ligand Activity

A protein of the present invention may also demonstrate
activity as receptors, receptor ligands or inhibitors or agonists
of receptor/ligand interactions. Examples of such receptors and
ligands include, without limitation, cytokine receptors and their
30 ligands, receptor kinases and their ligands, receptor
phosphatases and their ligands, receptors involved in cell-cell
interactions and their ligands (including without limitation,
cellular adhesion molecules (such as selectins, integrins and
their ligands) and receptor/ligand pairs involved in antigen
35 presentation, antigen recognition and development of cellular and
humoral immune responses). Receptors and ligands are also useful
for screening of potential peptide or small molecule inhibitors

of the relevant receptor/ligand interaction. A protein of the present invention (including, without limitation, fragments of receptors and ligands) may themselves be useful as inhibitors of receptor/ligand interactions.

5 The activity of a protein of the invention may, among other means, be measured by the following methods:

 Suitable assays for receptor-ligand activity include without limitation those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M.
10 Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 7.28, Measurement of Cellular Adhesion under static conditions 7.28.1-7.28.22), Takai et al., Proc. Natl. Acad. Sci. USA 84:6864-6868, 1987; Bierer et al., J. Exp. Med. 168:1145-1156, 1988; Rosenstein et al., J. Exp. Med.
15 169:149-160 1989; Stoltenborg et al., J. Immunol. Methods 175:59-68, 1994; Stitt et al., Cell 80:661-670, 1995.

Anti-Inflammatory Activity

 Proteins of the present invention may also exhibit anti-inflammatory activity. The anti-inflammatory activity may be
20 achieved by providing a stimulus to cells involved in the inflammatory response, by inhibiting or promoting cell-cell interactions (such as, for example, cell adhesion), by inhibiting or promoting chemotaxis of cells involved in the inflammatory process, inhibiting or promoting cell extravasation, or by
25 stimulating or suppressing production of other factors which more directly inhibit or promote an inflammatory response. Proteins exhibiting such activities can be used to treat inflammatory conditions including chronic or acute conditions), including without limitation intimation associated with infection (such as
30 septic shock, sepsis or systemic inflammatory response syndrome (SIRS)), ischemia-reperfusion injury, endotoxin lethality, arthritis, complement-mediated hyperacute rejection, nephritis, cytokine or chemokine-induced lung injury, inflammatory bowel disease, Crohn's disease or resulting from over production of
35 cytokines such as TNF or IL-1. Proteins of the invention may also be useful to treat anaphylaxis and hypersensitivity to an antigenic substance or material.

Tumor Inhibition Activity

In addition to the activities described above for immunological treatment or prevention of tumors, a protein of the invention may exhibit other anti-tumor activities. A protein may
5 inhibit tumor growth directly or indirectly (such as, for example, via ADCC). A protein may exhibit its tumor inhibitory activity by acting on tumor tissue or tumor precursor tissue, by inhibiting formation of tissues necessary to support tumor growth
10 (such as, for example, by inhibiting angiogenesis), by causing production of other factors, agents or cell types which inhibit tumor growth, or by suppressing, eliminating or inhibiting factors, agents or cell types which promote tumor growth.

Other Activities

A protein of the invention may also exhibit one or more of
15 the following additional activities or effects: inhibiting the growth, infection or function of, or killing, infectious agents, including, without limitation, bacteria, viruses, fungi and other parasites; effecting (suppressing or enhancing) bodily characteristics, including, without limitation, height, weight,
20 hair color, eye color, skin, fat to lean ratio or other tissue pigmentation, or organ or body part size or shape (such as, for example, breast augmentation or diminution, change in bone form or shape); effecting biorhythms or circadian cycles or rhythms; effecting the fertility of male or female subjects; effecting the
25 metabolism, catabolism, anabolism, processing, utilization, storage or elimination of dietary fat, lipid, protein, carbohydrate, vitamins, minerals, cofactors or other nutritional factors or component(s); effecting behavioral characteristics, including, without limitation, appetite, libido, stress,
30 cognition (including cognitive disorders), depression (including depressive disorders) and violent behaviors; providing analgesic effects or other pain reducing effects; promoting differentiation and growth of embryonic stem cells in lineages other than hematopoietic lineages; hormonal or endocrine activity; in the
35 case of enzymes, correcting deficiencies of the enzyme and treating deficiency-related diseases; treatment of hyperproliferative disorders (such as, for example, psoriasis); immunoglobulin-like activity (such as, for example, the ability

to bind antigens or complement)); and the ability to act as an antigen in a vaccine composition to raise an immune response against such protein or another material or entity which is cross-reactive with such protein.

5 Particular Applications for Certain Clones

The following sets out a non-exclusive list of applications for certain embodiments of the invention. In the interest of economy, applications relevant to multiple embodiments are not duplicated in this list. Other embodiments described herein have similar characteristics, as described there. The artisan is directed, therefore, to the Description of the Sequences for similar descriptions of the functions of other embodiment.

Testes

15 htes3_10i1b: The new protein can find application in diagnosis/therapy in leukemia predisposition/disease in the modulation of DNA repair.

20 htes3_10n10: The new protein can find application in studying the expression profile of testis-specific genes.

htes3_11a17: The new protein can find application in studying the expression profile of testis-specific genes and as a new marker for testicular cells.

25 htes3_11c22: The new protein can find application in modulating/blocking of regulatory pathways.

30 htes3_11d21: The new protein can find application in diagnosis of diseases due to unnormal protein degradation like muscular dystrophy or multiple sclerosis as well as in modulating the half life of specific proteins and in expression profiling.

35 Kidney

hfkd2_3k1 The new protein can find application in modulation of endocytosis. strong similarity to testicular dynamin (Rattus norvegicus).

Amygdala:

hamy2_10h17: The new protein can find application in
modulating protein-protein-interaction and in studying the
expression profile of amygdala-specific genes.

hamy2_10p7: The new protein can find application in
modulation of Na^+/Ca^{2+} -exchange and voltage-dependend
processes.

hamy2_11d2: The new protein can find application in studying
the expression profile of amygdala-specific genes and as a
new marker for amygdala cells.

hamy2_11n4: The new protein can find application in
modulation of DNA-repair and a as a new tool for
manipulation of nucleic acids.

hamy2_121f19: The new protein can find application
modulation of cyto skeleton-membrane interactions.

Fetal Brain:

hfbr2_78c12: The new protein can find application in the
modulation of translational pathways.

hfbr2_78d18: The new protein can find application in
studying the expression profile of brain-specific genes.

hfbr2_78d4: The new protein can find application in studying
the expression profile of brain-specific genes and as a new
marker for amygdala cells.

hfbr2_78e18: The new protein can find application in
studying the expression profile of brain-specific genes.

hfbr2_78i21: The new protein can find application in
diagnosis/modulation of protein damage and age-related
degenerative processes.

Melanoma:

hmel2_12j1: The new protein can find application in studying the expression profile of melanoma-specific genes.

hmel2_7g14: The new protein can find application in modulation of the sorting of proteins into different compartments.

hmel2_7k19: The new protein can find application in studying the expression profile of melanoma-specific genes.

VARIANTS OF THE INVENTIVE DNA MOLECULES

Variants in General

"Variants," according to the invention, include DNA and/or protein molecules that resemble, structurally and/or functionally, those set forth herein. Variants may be isolated from natural sources ("homologs"), may be entirely synthetic or may be based in part on both natural and synthetic approaches.

The section set forth below presents various structural and functional characteristics of molecules within the invention. Preferred molecules are characterized by a combination of one or more of these characteristics. For instance, some preferred molecules are described with reference to at least two structural characteristics, while others may be described with reference to at least one structural and at least one functional characteristic.

It will be recognized by the skilled artisan that structure ultimately defines function, i.e. the functions of the molecules described herein derives from the structures of those molecules. Accordingly, the structural variants described below that bear the closest structural relationship (as variously defined below) to the inventive molecules are the variants that most likely will preserve biological function. This relationship between structure and function will guide the skilled artisan in identifying the preferred embodiments of the invention.

Splicing Variants

It is well-known that eukaryotic structural genes are comprised of both protein coding and non-coding portions. When the messenger RNA is transcribed from the DNA template, it contains introns, which are non-coding, and exons, which are coding. In order to form a translation competent mRNA, the introns must be "spliced" out of this initial pre mRNA.

Specific sequences within the pre mRNA represent "splice junctions" that direct the cellular splicing machinery to the appropriate position. The splice junctions are loosely conserved sequence regions of the pre mRNA, which almost invariably begin with GT and end with AG (DNA perspective). The 5' end of the splice junction typically contains about nine somewhat conserved residues, for example, C/AAGTA/GAGT. The 3' end usually contains a pyrimidine rich stretch of at least about 11 nucleotides, followed by NC/TAGG. Splicing occurs before the GT and after the AG. Mount, *Nucleic Acids Res.* 10:459-72 (1982).

Interestingly, exons often correspond to discrete functional domains of the protein product. The intron/exon arrangement thus creates a linear array of nucleotides which can be correlated to discrete, and often interchangeable, functional protein fragments. Go, *Nature* 291:90-92 (1981); Branden et al., *EMBO J.* 3:1307-10 (1984). This linear arrangement creates the possibility of generating multiple different full length proteins by rearranging the order of the different functional portions in the array. For example, if a set of exons are arranged 1-2-3-4, where (-) represents the introns separating the exons, a splicing event need not simply produce 1234, but may produce 123, 134, 124 and so on. Production of different mRNA products in this way is commonly called "alternative splicing." Andreadis et al., *Ann. Rev. Cell Biol.* 3:207-42 (1987).

Some of the present DNA molecules can be represented in modular fashion in terms of their coding regions. Essentially, these modules are exons (though each "exon" may in fact be made up of several exons), which may be combined in different ways to form a variety of different DNA molecules, each encoding a different functional protein. Splicing variants are indicated in the Description of the Sequences.

Degenerate Variants

One aspect of the present invention provides "degenerate variants" of the nucleic acid fragments of the present invention. A "degenerate variant" is a nucleotide fragment which differs from those of inventive molecules by nucleotide sequence, but due to the degeneracy of the genetic code, encodes an identical polypeptide sequence.

Given the known relationship between DNA sequences and the proteins they encode, degenerate variants typically are described by reference to this relationship. It is well known that the degeneracy of the genetic code results in many possible DNA sequences which encode a particular protein. Indeed, of the three bases which comprise an amino acid-encoding triplet, the third position, and often the second, almost always may vary. This fact alone allows for a class of variant DNA molecules which encode protein sequences identical to those disclosed herein, yet have about 30% sequence variation. In other words, the variant DNA molecules are about 70% identical to the inventive DNAs, having no additional or deleted sequences. Thus, one aspect of the invention provides degenerate variant DNA molecules encoding the inventive protein sequences.

In one embodiment, these variants have at least about 70% sequence identity with the DNA molecules described herein. In a preferred embodiment, these variants have at least about 80% sequence identity to the inventive molecules. In a more preferred embodiment these variants have at least about 90% sequence identity with the inventive molecules.

Conservative Amino Acid Variants

Variants according to the invention also may be made that conserve the overall molecular structure of the encoded proteins. Given the properties of the individual amino acids comprising the disclosed protein products, some rational substitutions will be recognized by the skilled worker. Amino acid substitutions, i.e. "conservative substitutions," may be made, for instance, on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues involved.

For example: (a) nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine; (b) polar neutral amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine; (c) positively charged (basic) amino acids include arginine, lysine, and histidine; and (d) negatively charged (acidic) amino acids include aspartic acid and glutamic acid. Substitutions typically may be made within groups (a)-(d). In addition, glycine and proline may be substituted for one another based on their ability to disrupt α -helices. Similarly, certain amino acids, such as alanine, cysteine, leucine, methionine, glutamic acid, glutamine, histidine and lysine are more commonly found in α -helices, while valine, isoleucine, phenylalanine, tyrosine, tryptophan and threonine are more commonly found in β -pleated sheets. Glycine, serine, aspartic acid, asparagine, and proline are commonly found in turns. Some preferred substitutions may be made among the following groups: (i) S and T; (ii) P and G; and (iii) A, V, L and I. Given the known genetic code, and recombinant and synthetic DNA techniques, the skilled scientist readily can construct DNAs encoding the conservative amino acid variants.

As used herein, "sequence identity" between two polypeptide sequences indicates the percentage of amino acids that are identical between the sequences. "Sequence similarity" indicates the percentage of amino acids that either are identical or that represent conservative amino acid substitutions.

Functionally Equivalent Variants

Yet another class of DNA variants within the scope of the invention may be described with reference to the product they encode. As shown in the Description of the Sequences, some of the inventive DNA molecules encode a protein having a degree of homology with known proteins, or protein domains. It is expected, therefore, that they will have some or all of the requisite functional features of such molecules. These "functionally equivalent variants" products are characterized by the fact that they are functionally equivalent, with respect to biological activity, to certain known molecules.

Also provided herein is information on common structural motifs, including consensus sequences that will guide the artisan in constructing functionally equivalent variants. It will be understood that the motifs, identified in the Description of the Sequences for each inventive protein, may be modified within the identified consensus sequences. Thus, the invention contemplates the proteins in the Description of the Sequences that contain variability in the consensus sequences identified, and the invention further contemplates the full range of nucleic acids encoding them, and the complements of those nucleic acids.

Hybridizing Variants

DNA variants within the invention also may be described by reference to their physical properties in hybridization. One skilled in the field will recognize that DNA can be used to identify its complement and, since DNA is double stranded, its equivalent or homolog, using nucleic acid hybridization techniques. It will also be recognized that hybridization can occur with less than 100% complementarity. However, given appropriate choice of conditions, hybridization techniques can be used to differentiate among DNA sequences based on their structural relatedness to a particular probe. For guidance regarding such conditions see, for example, Sambrook et al., 1989, MOLECULAR CLONING, A LABORATORY MANUAL, Cold Spring Harbor Press, N.Y.; and Ausubel et al., 1989, CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, Green Publishing Associates and Wiley Interscience, N.Y.

Structural relatedness between two polynucleotide sequences can be expressed as a function of "stringency" of the conditions under which the two sequences will hybridize with one another. As used herein, the term "stringency" refers to the extent that the conditions disfavor hybridization. Stringent conditions strongly disfavor hybridization, and only the most structurally related molecules will hybridize to one another under such conditions. Conversely, non-stringent conditions favor hybridization of molecules displaying a lesser degree of structural relatedness. Hybridization stringency, therefore, directly correlates with the structural relationships of two nucleic acid sequences. The following relationships are useful in correlating hybridization

and relatedness (where T_m is the melting temperature of a nucleic acid duplex):

5 a. $T_m = 69.3 + 0.41(G+C)\%$

 b. The T_m of a duplex DNA decreases by 1°C with every increase of 1% in the number of mismatched base pairs.

10 c. $(T_m)_{\mu 2} - (T_m)_{\mu 1} = 18.5 \log_{10} \mu 2 / \mu 1$

 where $\mu 1$ and $\mu 2$ are the ionic strengths of two solutions.

 Hybridization stringency is a function of many factors, including overall DNA concentration, ionic strength, temperature, 15 probe size and the presence of agents which disrupt hydrogen bonding. Factors promoting hybridization include high DNA concentrations, high ionic strengths, low temperatures, longer probe size and the absence of agents that disrupt hydrogen bonding.

20 Hybridization usually is done in two stages. First, in the "binding" stage, the probe is bound to the target under conditions favoring hybridization. Stringency is usually controlled at this stage by altering the temperature. For high stringency, the temperature is usually between 65°C and 70°C , unless short (<20 25 nt) oligonucleotide probes are used. A representative hybridization solution comprises 6X SSC, 0.5% SDS, 5X Denhardt's solution and 100 μg of non-specific carrier DNA. See Ausubel et al., *supra*, section 2.9, supplement 27 (1994). Of course many different, yet functionally equivalent, buffer conditions are 30 known. Where the degree of relatedness is lower, a lower temperature may be chosen. Low stringency binding temperatures are between about 25°C and 40°C . Medium stringency is between at least about 40°C to less than about 65°C . High stringency is at least about 65°C .

35 Second, the excess probe is removed by washing. It is at this stage that more stringent conditions usually are applied. Hence, it is this "washing" stage that is most important in determining relatedness via hybridization. Washing solutions typically contain lower salt concentrations. One exemplary medium 40 stringency solution contains 2X SSC and 0.1% SDS. A high stringency wash solution contains the equivalent (in ionic

strength) of less than about 0.2X SSC, with a preferred stringent solution containing about 0.1X SSC. The temperatures associated with various stringencies are the same as discussed above for "binding." The washing solution also typically is replaced a number of times during washing. For example, typical high stringency washing conditions comprise washing twice for 30 minutes at 55° C. and three times for 15 minutes at 60° C.

The present invention includes nucleic acid molecules that hybridize to the inventive molecules under high stringency binding and washing conditions. More preferred molecules (from an mRNA perspective) are those that are at least 50 % of the length of any one of those depicted in the Description of the Sequences. Particularly preferred molecules are at least 75 % of the length of those molecules.

15 *Substitutions, Insertions, Additions and Deletions*

In a general sense, the preferred DNA variants of the invention are those that retain the closest relationship, as described by "sequence identity" to the inventive DNA molecules. According to another aspect of the invention, therefore, substitutions, insertions, additions and deletions of defined properties are contemplated. It will be recognized that sequence identity between two polynucleotide sequences, as defined herein, generally is determined with reference to the protein coding region of the sequences. Thus, this definition does not at all limit the amount of DNA, such as vector DNA, that may be attached to the molecules described herein. Preferred DNA sequence variants include molecules encoding proteins sharing some or all of any relevant biological activity of the native molecule.

In creating these variants, the skilled worker will be guided by reference to the protein structure. First, insertions and deletions in any recognized functional domain above generally should be avoided, except as noted below in the section entitled "Proteins," where this domain is discussed in detail. Alterations in such domains usually will be limited to conservative amino acid substitutions. In addition, where insertions and deletions are desired, this may be accomplished at the N- and/or C-terminus of the protein molecule (or the corresponding coding regions of the DNA). If insertions or deletions are made within the protein,

deletions of major structural features usually should be avoided. Thus, a preferred place to make insertion or deletion variants is in non-structural regions, such as linker regions between two alpha helices.

5 "Substitutions" generally refer to alterations in the DNA sequence which do not change its overall length, but only alter one or more nucleotide positions, substituting one for another in the common sense of the word. One class of preferred substitutions, "degenerate substitutions," are those that do not
10 alter the encoded amino acid sequence. Some substitutions retain 50%, 55%, 60% or 65% identity. Preferred substitutions retain at least about 70% identity, more preferably at least 70% or 75% identity, with the inventive DNAs. Some more preferred molecules have at least about 80% identity, more preferably at least 80% or
15 85% identity. Particularly preferred DNAs share at least about 90% identity, more preferably at least 90% or 95% identity.

"Insertions," unlike substitutions, alter the overall length of the DNA molecule, and thus sometimes the encoded protein. Insertions add extra nucleotides to the interior (not the 5' or 3'
20 ends) of the subject DNAs. Preferred insertions are made with reference to the protein sequence encoded by the DNA. Thus, it is most preferred to provide an insertion in the DNA at a location that corresponds to an area of the encoded protein which lacks structure. For instance, it typically would not be beneficial, if
25 the preservation of biological activity is desired, to provide an insertion within an alpha-helical region or a beta-pleated sheet. Accordingly, non-structural areas, such as those containing helix-breaking glycines and proline residues, are most preferred sites of insertion. Other preferred sites of insertion are the splice
30 sites, which are indicated above in the description of the inventive DNA molecules.

While the optimal size of insertions will vary depending upon the site of insertion and its effect on the overall conformation of the encoded protein, some general guides are useful.
35 Generally, the total insertions (irrespective of their number) should not add more than about 30% (or preferably not more than 30%) to the overall size of the encoded protein. More preferably, the insertion adds less than about 10-20% (yet more preferably 10-20%) in size, with less than about 10% being most preferred. The

number of insertions is limited only by the number of suitable insertions sites, and secondarily by the foregoing size preferences.

"Additions," like insertions, also add to the overall size of the DNA molecule, and usually the encoded protein. However, instead of being made within the molecule, they are made on the 5' or 3' end, usually corresponding to the N- or C- terminus of the encoded protein. Unlike deletions, additions are not very size-dependent. Indeed, additions may be of virtually any size. Preferred additions, however, do not exceed about 100% of the size of the native molecule. More preferably, they add less than about 60 to 30% to the overall size, with less than about 30% being most preferred.

"Deletions" diminish the overall size of the DNA and, therefore, also reduce the size of the protein encoded by that DNA. Deletions may be made from either end of the molecule or internal to it. Typical preferred deletions remove discrete structural features of the encoded protein. For example, some deletions will comprise the deletion of one or more exons which may define a structural feature. Preferred deletions remove less than about 30% of the size of the subject molecule. More preferred deletions remove less than about 20% and most preferred deletions remove less than about 10%.

Computer-Defined Variants and Definition of "Sequence Identity"

In general, both the DNA and protein molecules of the invention can be defined with reference to "sequence identity." As used herein, "sequence identity" refers to a comparison made between two molecules using, for example, the standard Smith-Waterman algorithm that is well known in the art.

Some molecules have at least about 50%, 55% or 60% identity. Preferred molecules are those having at least about 65% sequence identity, more preferably at least 65% or 70% sequence identity. Other preferred molecules have at least about 80%, more preferably at least 80% or 85%, sequence identity. Particularly preferred molecules have at least about 90% sequence identity, more preferably at least 90% sequence identity. Most preferred molecules have at least about 95%, more preferably at least 95%, sequence identity. As used herein, two nucleic acid molecules or

proteins are said to "share significant sequence identity" if the two contain regions which possess greater than 85% sequence (amino acid or nucleic acid) identity.

"Sequence identity" is defined herein with reference the

- 5 Blast 2 algorithm, which is available at the NCBI
(<http://www.ncbi.nlm.nih.gov/BLAST>), using default parameters.
References pertaining to this algorithm include: those found at
http://www.ncbi.nlm.nih.gov/BLAST/blast_references.html;
Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J.
10 (1990) "Basic local alignment search tool." J. Mol. Biol.
215:403-410; Gish, W. & States, D.J. (1993) "Identification of
protein coding regions by database similarity search." Nature
Genet. 3:266-272; Madden, T.L., Tatusov, R.L. & Zhang, J. (1996)
"Applications of network BLAST server" Meth. Enzymol. 266:131-
15 141; Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.,
Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and
PSI-BLAST: a new generation of protein database search programs."
Nucleic Acids Res. 25:3389-3402; and Zhang, J. & Madden, T.L.
(1997) "PowerBLAST: A new network BLAST application for
20 interactive or automated sequence analysis and annotation."
Genome Res. 7:649-656.

METHODS OF MAKING VARIANTS

It will be recognized that variants of the inventive
molecules can be constructed in several different ways. For
25 example, they may be constructed as completely synthetic DNAs.
Methods of efficiently synthesizing oligonucleotides in the range
of 20 to about 150 nucleotides are widely available. See Ausubel
et al., *supra*, section 2.11, Supplement 21 (1993). Overlapping
oligonucleotides may be synthesized and assembled in a fashion
30 first reported by Khorana *et al.*, J. Mol. Biol. 72:209-217 (1971);
see also Ausubel *et al.*, Section 8.2. The synthetic DNAs are
designed with convenient restriction sites engineered at the 5'
and 3' ends of the gene to facilitate cloning into an appropriate
vector.

35 An alternative method of generating variants is to start with
one of the inventive DNAs and then to conduct site-directed
mutagenesis. See Ausubel *et al.*, *supra*, chapter 8, Supplement 37

(1997). In a typical method, a target DNA is cloned into a single-stranded DNA bacteriophage vehicle. Single-stranded DNA is isolated and hybridized with a oligonucleotide containing the desired nucleotide alteration(s). The complementary strand is synthesized and the double stranded phage is introduced into a host. Some of the resulting progeny will contain the desired mutant, which can be confirmed using DNA sequencing. In addition, various methods are available that increase the probability that the progeny phage will be the desired mutant. These methods are well known to those in the field and kits are commercially available for generating such mutants.

ISOLATING HOMOLOGS

Methods

By using the sequences disclosed herein as probes or as primers, and techniques such as PCR cloning and colony/plaque hybridization, one skilled in the art can obtain homologs. "Homologs" are essentially naturally-occurring variants and include allelic, species-specific and tissue-specific variants.

Region-specific primers or probes derived from the nucleotide sequence(s) provided can be used to prime DNA synthesis and PCR amplification, as well as to identify colonies containing cloned DNA encoding a homolog using known methods (Innis et al., *PCR Protocols*, Academic Press, San Diego, CA (1990)). Such an application is useful in diagnostic methods, as described in more detail below, as well as in preparing full-length DNAs from various sources. The PCR primers are preferably at least 15 bases, and more preferably at least 18 bases in length. When selecting a primer sequence, it is preferred that the primer pairs have approximately the same G/C ratio, so that melting temperatures are approximately the same. As a general guide, the formula $3(G+C) + 2(A+T) = ^\circ C$, is useful.

When using primers derived from the inventive sequences, one skilled in the art will recognize that by employing high stringency conditions (e.g., annealing at 50-60°C), only sequences with greater than 75% sequence identity to the primer will be amplified. By employing lower stringency conditions (e.g.,

annealing at 35-37°C), sequences which have greater than 40-50% sequence identity to the primer also will be amplified.

The PCR product may be subcloned and sequenced to confirm that it indeed displays the expected sequence identity. The PCR
5 fragment may then be used to isolate a full length cDNA clone by a variety of methods. For example, the amplified fragment may be labeled and used to screen a bacteriophage cDNA library. Alternatively, the labeled fragment may be used to screen a genomic library.

10 PCR technology may also be utilized to isolate full length cDNA sequences. For example, RNA may be isolated, following standard procedures, from an appropriate cellular or tissue source. A reverse transcription reaction may be performed on the RNA using an oligonucleotide primer specific for the most 5' end
15 of the amplified fragment for the priming of first strand synthesis. The resulting RNA/DNA hybrid may then be "tailed" with guanines using a standard terminal transferase reaction, the hybrid may be digested with RNAase H, and second strand synthesis may then be primed with a poly-C primer. Thus, cDNA sequences
20 upstream of the amplified fragment may easily be isolated. For a review of cloning strategies which may be used, see e.g., Sambrook et al., 1989, *supra*.

When using DNA probes derived from the inventive sequences for colony/plaque hybridization, one skilled in the art will
25 recognize that by employing medium to high stringency conditions (e.g., hybridizing at 50-65°C in 5X SSPE and 50% formamide, and washing at 50-65°C in 0.5X SSPE), sequences having regions with greater than 70% sequence identity to the probe can be obtained, and that by employing lower stringency conditions (e.g.,
30 hybridizing at 35-37°C in 5X SSPE and 40-45% formamide, and washing at 42°C in SSPE), sequences having regions with greater than 35-45% sequence identity to the probe will be obtained.

Suitably, genomic or cDNA libraries can be constructed and screened in accord with the previous paragraph. The libraries
35 should be derived from a tissue or organism that is known to express the gene of interest, or that is suspected of expressing the gene. The clone containing the homolog may then be purified

through methods routinely practiced in the art, and subjected to sequence analysis.

5 Additionally, an expression library can be constructed utilizing DNA isolated from or cDNA synthesized from a tissue or organism that is known to express the gene of interest, or that is suspected of expressing the gene. In this manner, clones may be induced and screened using standard antibody screening techniques in conjunction with antibodies raised against the normal gene product, as described herein. (For screening techniques, see, for example, Harlow, E. and Lane, eds., 1988, ANTIBODIES: A
10 LABORATORY MANUAL, Cold Spring Harbor Press, Cold Spring Harbor Press.)

Human Homologs

Any organism or tissue can be used as the source for homologs
15 of the present invention so long as the organism or tissue naturally expresses such a protein or contains genes encoding the same. The most preferred organism for isolating homologs is human.

PROTEINS OF THE INVENTION

20 One class of proteins included within the invention is encoded by the inventive DNA molecules presented. Other proteins according to the invention are those encoded by the DNA variants described above. As noted, these variants are designed with the encoded proteins in mind.

25 A preferred class of protein fragments includes those fragments which retain any biological activity. These molecules share functional features common the family of proteins, although these characteristics may vary in degree.

According to one aspect of the invention fragments of the
30 inventive proteins are contemplated. Some preferred fragments are those which are capable of eliciting an immune response. Generally these "antigenic" fragments will be from about five amino acids in length to about fifty amino acids in length. Some preferred antigenic fragments are from five to about twenty amino
35 acids long. "Antigenic" response may refer to a T cell response, a B cell response or a response by cells of the macrophage/monocyte lineages. In most cases, however, it will

refer to the immune response involved in the generation of antibodies. In other words, the relevant immune response is that of helper T cells and/or B cells. These preferred molecules comprise one or more T cell and /or B cell epitopes.

5 ANTIBODIES OF THE INVENTION

Antibodies raised against the proteins and protein fragments of the invention also are contemplated by the invention. Described below are antibody products and methods for producing antibodies capable of specifically recognizing one or more
10 epitopes of the presently described proteins and their derivatives.

Antibodies include, but are not limited to polyclonal antibodies, monoclonal antibodies (mAbs), humanized or chimeric antibodies, single chain antibodies including single chain Fv
15 (scFv) fragments, Fab fragments, F(ab')₂ fragments, fragments produced by a Fab expression library, anti-idiotypic (anti-Id) antibodies, epitope-binding fragments, and humanized forms of any of the above.

As known to one in the art, these antibodies may be used, for
20 example, in the detection of a target protein in a biological sample. They also may be utilized as part of treatment methods, and/or may be used as part of diagnostic techniques whereby patients may be tested for abnormal levels or for the presence of abnormal forms of the such proteins.

In general, techniques for preparing polyclonal and
25 monoclonal antibodies as well as hybridomas capable of producing the desired antibody are well known in the art (Campbell, A.M., *Monoclonal Antibody Technology: Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Science Publishers, Amsterdam, The Netherlands (1984); St. Groth et al., *J. Immunol. Methods* 35:1-21 (1980); Kohler and Milstein, *Nature* 256:495-497 (1975)), the trioma technique, the human B-cell hybridoma technique (Kozbor et al., *Immunology Today* 4:72 (1983); Cole et al., in *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, Inc. (1985), pp. 77-96). Antibodies may also be generated by the
30 known techniques of phage display and *in vitro* immunization.

Polyclonal Antibodies

Polyclonal antibodies are heterogeneous populations of antibody molecules derived from the sera of animals immunized with an antigen, such as an inventive protein or an antigenic derivative thereof.

Polyclonal antiserum, containing antibodies to heterogeneous epitopes of a single protein, can be prepared by immunizing suitable animals with the expressed protein described above, which can be unmodified or modified, as known in the art, to enhance immunogenicity. Immunization methods include subcutaneous or intraperitoneal injection of the polypeptide.

Effective polyclonal antibody production is affected by many factors related both to the antigen and to the host species. For example, small molecules tend to be less immunogenic than other and may require the use of carriers and/or adjuvant. In addition, host animal response may vary with site of inoculation. Both inadequate or excessive doses of antigen may result in low titer antisera. In general, however, small doses (high ng to low μ g levels) of antigen administered at multiple intradermal sites appears to be most reliable. Host animals may include but are not limited to rabbits, mice, chickens and rats, to name but a few. An effective immunization protocol for rabbits can be found in Vaitukaitis, J. et al., *J. Clin. Endocrinol. Metab.* 33:988-991 (1971).

The protein immunogen may be modified or administered in an adjuvant in order to increase the protein's antigenicity. Methods of increasing the antigenicity of a protein are well known in the art and include, but are not limited to coupling the antigen with a heterologous protein (such as globulin β -galactosidase) or through the inclusion of an adjuvant during immunization. Adjuvants include Freund's (complete and incomplete), mineral gels such as aluminum hydroxide, surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanin, dinitrophenol, and potentially useful human adjuvants such as BCG (bacille Calmette-Guerin) and *Corynebacterium parvum*.

Booster injections can be given at regular intervals, with at least one usually being required for optimal antibody production.

The antiserum may be harvested when the antibody titer begins to fall. Titer may be determined semi-quantitatively, for example, by double immunodiffusion in agar against known concentrations of the antigen. See, for example, Ouchterlony et al., Chap. 19 in:
5 *Handbook of Experimental Immunology*, Wier, ed. Blackwell (1973).
Plateau concentration of antibody is usually in the range of 0.1 to 0.2 mg/ml of serum (about 12 μ M). The antiserum may be purified by affinity chromatography using the immobilized immunogen carried on a solid support. Such methods of affinity
10 chromatography are well known in the art.

Affinity of the antisera for the antigen may be determined by preparing competitive binding curves, as described, for example, by Fisher, Chap. 42 in: *Manual of Clinical Immunology*, second edition, Rose and Friedman, eds., Amer. Soc. For Microbiology,
15 Washington, D.C. (1980).

In addition to using protein as the immunogen, DNA molecules may be used directly. In this manner, a DNA encoding the protein immunogen is administered. Boosting and harvesting is done in a manner analogous to that detailed above. Yet another method of
20 producing antibodies entails immunizing chickens and harvesting the antibodies from their eggs.

Monoclonal Antibodies

Monoclonal antibodies (MAbs), are homogeneous populations of antibodies to a particular antigen. They may be obtained by any
25 technique which provides for the production of antibody molecules by continuous cell lines in culture or *in vivo*. MAbs may be produced by making hybridomas which are immortalized cells capable of secreting a specific monoclonal antibody.

Monoclonal antibodies to any of the proteins, peptides and epitopes thereof described herein can be prepared from murine hybridomas according to the classical method of Kohler, G. and Milstein, C., *Nature* 256:495-497 (1975) (and U.S. Patent No. 4,376,110) or modifications of the methods thereof, such as the human B-cell hybridoma technique (Kosbor et al., 1983, *Immunology Today* 4:72; Cole et al., 1983, *Proc. Natl. Acad. Sci. USA* 80:2026-2030), and the EBV-hybridoma technique (Cole et al., 1985, *MONOCLONAL ANTIBODIES AND CANCER THERAPY*, Alan R. Liss, Inc., pp. 77-96).

In one method a mouse is repetitively inoculated with a few micrograms of the selected protein over a period of a few weeks. The mouse is then sacrificed, and the antibody producing cells of the spleen are isolated.

The spleen cells are fused, typically using polyethylene glycol, with mouse myeloma cells, such as SP2/O-Ag14 myeloma cells. The excess, unfused cells are destroyed by growth of the system on selective media comprising aminopterin (HAT media). The successfully fused cells are diluted, and aliquots are plated to microliter plates where growth is continued. Antibody--producing clones (hybridomas) are identified by detection of antibody in the supernatant fluid of the wells by immunoassay procedures. These include ELISA, as originally described by Engvall, *Meth. Enzymol.* 70:419 (1980), western blot analysis, radioimmunoassay (Lutz et al., *Exp. Cell Res.* 175:109-124 (1988)) and modified methods thereof.

Selected positive clones can be expanded and their monoclonal antibody product harvested for use. Detailed procedures for monoclonal antibody production are described in Davis, L. et al. *BASIC METHODS IN MOLECULAR BIOLOGY*, Elsevier, New York. Section 21-2 (1989). The hybridoma clones may be cultivated *in vitro* or *in vivo*, for instance as ascites. Production of high titers of mAbs *in vivo* makes this the presently preferred method of production. Alternatively, hybridoma culture in hollow fiber bioreactors provides a continuous high yield source of monoclonal antibodies.

The antibody class and subclass may be determined using procedures known in the art (Campbell, A.M., *Monoclonal Antibody*

Technology: *Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Science Publishers, Amsterdam, The Netherlands (1984)). MAbs may be of any immunoglobulin class including IgG, IgM, IgE, IgA, IgD and any subclass thereof. Methods of purifying
5 monoclonal antibodies are well known in the art.

Antibody Derivatives and Fragments

Fragments or derivatives of antibodies include any portion of the antibody which is capable of binding the target antigen, or a specific portion thereof. Antibody derivatives include poly-
10 specific (e.g., bi-specific) antibodies, which contain binding sites specific for two or more different epitopes. These epitopes may be from the same or different inventive molecules or one or more epitope may be from a molecule not specifically disclosed here.

15 Antibody fragments specifically include $F(ab')_2$, Fab, Fab' and Fv fragments. These can be generated from any class of antibody, but typically are made from IgG or IgM. They may be made by conventional recombinant DNA techniques or, using the classical method, by proteolytic digestion with papain or pepsin.
20 See CURRENT PROTOCOLS IN IMMUNOLOGY, chapter 2, Coligan et al., eds., (John Wiley & Sons 1991-92).

$F(ab')_2$ fragments are typically about 110 kDa (IgG) or about 150 kDa (IgM) and contain two antigen-binding regions, joined at the hinge by disulfide bond(s). Virtually all, if not all, of the
25 Fc is absent in these fragments. Fab' fragments are typically about 55 kDa (IgG) or about 75 kDa (IgM) and can be formed, for example, by reducing the disulfide bond(s) of an $F(ab')_2$ fragment. The resulting free sulfhydryl group(s) may be used to conveniently conjugate Fab' fragments to other molecules, such as detection
30 reagents (e.g., enzymes).

Fab fragments are monovalent and usually are about 50 kDa (from any source). Fab fragments include the light (L) and heavy (H) chain, variable (V_L and V_H , respectively) and constant (C_L , C_H , respectively) regions of the antigen-binding portion of the
35 antibody. The H and L portions are linked by an intramolecular disulfide bridge.

Fv fragments are typically about 25 kDa (regardless of source) and contain the variable regions of both the light and

heavy chains (V_L and V_H , respectively). Usually, the V_L and V_H chains are held together only by non-covalent interactions and, thus, they readily dissociate. They do, however, have the advantage of small size and they retain the same binding properties of the larger Fab fragments. Accordingly, methods have been developed to crosslink the V_L and V_H chains, using, for example, glutaraldehyde (or other chemical crosslinkers), intermolecular disulfide bonds (by incorporation of cysteines) and peptide linkers. The resulting Fv is now a single chain (i.e., SCFv).

Other antibody derivatives include single chain antibodies (U.S. Patent 4,946,778; Bird, Science 242:423-426 (1988); Huston et al., Proc. Natl. Acad. Sci. USA 85:5879-5883 (1988); and Ward et al., Nature 334:544-546 (1989)). Single chain antibodies are formed by linking the heavy and light chain fragments of the Fv region via an amino acid bridge, resulting in a single chain Fv (SCFv).

One preferred method involves the generation of scFvs by recombinant methods, which allows the generation of Fvs with new specificities by mixing and matching variable chains from different antibody sources. In a typical method, a recombinant vector would be provided which comprises the appropriate regulatory elements driving expression of a cassette region. The cassette region would contain a DNA encoding a peptide linker, with convenient sites at both the 5' and 3' ends of the linker for generating fusion proteins. The DNA encoding a variable region(s) of interest may be cloned in the vector to form fusion proteins with the linker, thus generating an scFv.

In an exemplary alternative approach, DNAs encoding two Fvs may be ligated to the DNA encoding the linker, and the resulting tripartite fusion may be ligated directly into a conventional expression vector. The scFv DNAs generated any of these methods may be expressed in prokaryotic or eukaryotic cells, depending on the vector chosen.

Antibody fragments which recognize specific epitopes may be generated by known techniques. For example, such fragments include but are not limited to: the $F(ab')_2$ fragments which can be produced by pepsin digestion of the antibody molecule and the Fab fragments which can be generated by reducing the disulfide bridges

of the F(ab)₂ fragments. Alternatively, Fab expression libraries may be constructed (Huse et al., 1989, *Science*, 246:1275-1281) to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity.

5 Derivatives also include "chimeric antibodies" (Morrison et al., *Proc. Natl. Acad. Sci.*, 81:6851-6855 (1984); Neuberger et al., *Nature*, 312:604-608 (1984); Takeda et al., *Nature*, 314:452-454 (1985)). These chimeras are made by splicing the DNA encoding a mouse antibody molecule of appropriate specificity with, for
10 instance, DNA encoding a human antibody molecule of appropriate specificity. Thus, a chimeric antibody is a molecule in which different portions are derived from different animal species, such as those having a variable region derived from a murine mAb and a human immunoglobulin constant region. These are also known
15 sometimes as "humanized" antibodies and they offer the added advantage of at least partial shielding from the human immune system. They are, therefore, particularly useful in therapeutic *in vivo* applications.

Labeled Antibodies

20 The present invention further provides the above-described antibodies in detectably labeled form. Antibodies can be detectably labelled through the use of radioisotopes, affinity labels (such as biotin, avidin, etc.), enzymatic labels (such as horseradish peroxidase, alkaline phosphatase, etc.) fluorescent
25 labels (such as FITC or rhodamine, etc.), paramagnetic atoms, etc. Procedures for accomplishing such labeling are well-known in the art, for example see (Sternberger et al., *J. Histochem. Cytochem.* 18:315 (1970); Bayer et al., *Meth. Enzym.* 62:308 (1979); Engval et al., *Immunol.* 109:129 (1972); Goding, *J. Immunol. Meth.* 13:215
30 (1976)). The labeled antibodies of the present invention can be used for *in vitro*, *in vivo*, and *in situ* diagnostic assays.

Immobilized Antibodies

The foregoing antibodies also may be immobilized on a solid support. Examples of such solid supports include plastics such as polycarbonate, complex carbohydrates such as agarose and
35 sepharose, acrylic resins and such as polyacrylamide and latex beads. Techniques for coupling antibodies to such solid supports

are well known in the art (Weir et al., "Handbook of Experimental Immunology" 4th Ed., Blackwell Scientific Publications, Oxford, England, Chapter 10 (1986); Jacoby et al., Meth. Enzym. 34 Academic Press, N.Y. (1974)). The immobilized antibodies of the present invention can be used for *in vitro*, *in vivo*, and *in situ* assays as well as for immunoaffinity purification of the proteins of the present invention.

THERAPEUTIC AND DIAGNOSTIC COMPOSITIONS

The proteins, antibodies and polynucleotides of the present invention can be formulated according to known methods to prepare pharmaceutically useful compositions, whereby these materials, or their functional derivatives, are combined in admixture with a pharmaceutically acceptable carrier vehicle. Suitable vehicles and their formulation, inclusive of other human proteins, e.g., human serum albumin, are described, for example, in *Remington's Pharmaceutical Sciences* (16th ed., Osol, A., Ed., Mack, Easton PA (1980)). In order to form a pharmaceutically acceptable composition suitable for effective administration, such compositions will contain an effective amount of one or more of the agents of the present invention, together with a suitable amount of carrier vehicle.

Pharmaceutical compositions for use in accordance with the present invention may be formulated in conventional manner using one or more physiologically acceptable carriers or excipients. Thus, the compounds and their physiologically acceptable salts and solvate may be formulated for administration by inhalation or insufflation (either through the mouth or the nose) or oral, buccal, parenteral or rectal administration.

For oral administration, the pharmaceutical compositions may take the form of, for example, tablets or capsules prepared by conventional means with pharmaceutically acceptable excipients such as binding agents (e.g., pregelatinised maize starch, polyvinylpyrrolidone or hydroxypropyl methylcellulose); fillers (e.g., lactose, microcrystalline cellulose or calcium hydrogen phosphate); lubricants (e.g., magnesium stearate, talc or silica); disintegrants (e.g., potato starch or sodium starch glycolate); or

wetting agents (e.g., sodium lauryl sulphate). The tablets may be coated by methods well known in the art. Liquid preparations for oral administration may take the form of, for example, solutions, syrups or suspensions, or they may be presented as a dry product for constitution with water or other suitable vehicle before use. Such liquid preparations may be prepared by conventional means with pharmaceutically acceptable additives such as suspending agents (e.g., sorbitol syrup, cellulose derivatives or hydrogenated edible fats); emulsifying agents (e.g., lecithin or acacia); non-aqueous vehicles (e.g., almond oil, oily esters, ethyl alcohol or fractionated vegetable oils); and preservatives (e.g., methyl or propyl-p-hydroxybenzoates or sorbic acid). The preparations may also contain buffer salts, flavoring, coloring and sweetening agents as appropriate.

Preparations for oral administration may be suitably formulated to give controlled release of the active compound. For buccal administration the composition may take the form of tablets or lozenges formulated in conventional manner.

For administration by inhalation, the compounds for use according to the present invention are conveniently delivered in the form of an aerosol spray presentation from pressurized packs or a nebuliser, with the use of a suitable propellant, e.g., dichlorodifluoromethane, trichlorofluoromethane, dichlorotetrafluoroethane, carbon dioxide or other suitable gas.

In the case of a pressurized aerosol the dosage unit may be determined by providing a valve to deliver a metered amount. Capsules and cartridges of, e.g., gelatin for use in an inhaler or insufflator may be formulated containing a powder mix of the compound and a suitable powder base such as lactose or starch.

The compounds may be formulated for parenteral administration by injection, e.g., by bolus injection or continuous infusion. Formulations for injection may be presented in unit dosage form, e.g., in ampules or in multi-dose containers, with an added preservative. The compositions may take such forms as suspensions, solutions or emulsions in oily or aqueous vehicles, and may contain formulatory agents such as suspending, stabilizing and/or dispersing agents. Alternatively, the active ingredient

may be in powder form for constitution with a suitable vehicle, e.g., sterile pyrogen-free water, before use.

The compounds may also be formulated in rectal compositions such as suppositories or retention enemas, e.g., containing
5 conventional suppository bases such as cocoa butter or other glycerides.

In addition to the formulations described previously, the compounds may also be formulated as a depot preparation. Such long acting formulations may be administered by implantation (for
10 example subcutaneously or intramuscularly) or by intramuscular injection. Thus, for example, the compounds may be formulated with suitable polymeric or hydrophobic materials (for example as an emulsion in an acceptable oil) or ion exchange resins, or as sparingly soluble derivatives, for example, as a sparingly soluble
15 salt.

The compositions may, if desired, be presented in a pack or dispenser device which may contain one or more unit dosage forms containing the active ingredient. The pack may for example
comprise metal or plastic foil, such as a blister pack. The pack
20 or dispenser device may be accompanied by instructions for administration.

RECOMBINANT CONSTRUCTS AND EXPRESSION

The present invention further provides recombinant DNA constructs comprising one or more of the nucleotide sequences of
25 the present invention. The recombinant constructs of the present invention comprise a vector, such as a plasmid or viral vector, into which a DNA or DNA fragment, typically bearing an open reading frame, is inserted, in either orientation. The gene products encoded by the subject DNAs may be produced by
30 recombinant DNA technology using techniques well known in the art. See, for example, the techniques described in Sambrook et al., 1989, *supra*, and Ausubel et al., 1989, *supra*. Alternatively, the DNA sequences may be chemically synthesized using, for example, synthesizers. See, for example, the techniques described in
35 OLIGONUCLEOTIDE SYNTHESIS, 1984, Gait, ed., IRL Press, Oxford, which is incorporated by reference herein in its entirety. They may be assembled from fragments and short oligonucleotide linkers,

or from a series of oligonucleotides. They are preferably made by RT-PCR methods. The resulting synthetic gene is capable of being expressed in a recombinant vector.

In some cases the recombinant constructs will be expression vectors, which are capable of expressing the RNA and/or protein products of the encoded DNA(s). Thus, the vector may further comprise regulatory sequences, including for example, a promoter, operably linked to the open reading frame (ORF). The vector may further comprise a selectable marker sequence.

Specific initiation signals may also be required for efficient translation of inserted target gene coding sequences. These signals include the ATG initiation codon and adjacent sequences. In cases where a target DNA includes its own initiation codon and adjacent sequences is inserted into the appropriate expression vector, no additional translation control signals may be needed. However, in cases where only a portion of an ORF is used, exogenous translational control signals, including, perhaps, the ATG initiation codon, must be provided. Furthermore, the initiation codon must be in phase with the reading frame of the desired coding sequence to ensure translation of the entire target. These exogenous translational control signals and initiation codons can be of a variety of origins, both natural and synthetic. The efficiency of expression may be enhanced by the inclusion of appropriate transcription enhancer elements, transcription terminators, etc. (see Bittner et al., *Methods in Enzymol.* 153:516-544 (1987)). Some appropriate cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, et al., in *Molecular Cloning: A Laboratory Manual*, Second Edition, Cold Spring Harbor, New York (1989), the disclosure of which is hereby incorporated by reference.

If desired, to enhance expression and facilitate proper protein folding, the codon context and codon pairing of the sequence may be optimized for the particular expression organism, as explained by Hatfield et al., U.S. Patent No. 5,082,767.

The present invention further provides host cells containing at least one of the DNAs of the present invention. The host cell can be virtually any cell for which expression vectors are

available. It may be, for example, a higher eukaryotic host cell, such as a mammalian cell, a lower eukaryotic host cell, such as a yeast cell, or the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the recombinant construct into the host cell can be effected by calcium phosphate transfection, DEAE, dextran mediated transfection, or electroporation (Davis et al., *Basic Methods in Molecular Biology* (1986)).

A wide variety of expression systems are available, such as: yeast (e.g. *Saccharomyces*, *Pichia*) transformed with recombinant yeast expression vectors containing the target DNA; insect cell systems infected with recombinant virus expression vectors (e.g., baculovirus) containing the target DNA sequences; plant cell systems infected with recombinant virus expression vectors (e.g., cauliflower mosaic virus, CaMV; tobacco mosaic virus, TMV) or transformed with recombinant plasmid expression vectors (e.g. Ti plasmid) containing target DNA coding sequences; or mammalian cell systems (e.g. COS, CHO, BHK, 293, 3T3) harboring recombinant expression constructs containing promoters derived from the genome of mammalian cells (e.g., metallothionein promoter) or from mammalian viruses (e.g., the adenovirus late promoter; the vaccinia virus 7.5K promoter).

Depending on the system chosen, the resulting product may differ. For example, proteins expressed in most bacterial cultures, e.g., *E. coli*, will be free of glycosylation modifications; polypeptides or proteins expressed in yeast will have a glycosylation pattern different from that expressed in mammalian cells.

Vectors

Generally, recombinant expression vectors will include origins of replication and selectable markers permitting selection of the host cell, e.g., the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), α -factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate

phase with translation initiation and termination sequence, and in one aspect of the invention, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium. Optionally, the heterologous sequence can encode a fusion protein including an N-terminal or C-terminal identification peptide imparting desired characteristics, e.g., stabilization or simplified purification of expressed recombinant product.

Bacterial Expression

Useful expression vectors for bacterial use are constructed by inserting a structural DNA sequence encoding a desired protein together with suitable translation initiation and termination signals in operable reading phase with a functional promoter. The vector will comprise one or more phenotypic selectable markers and an origin of replication to ensure maintenance of the vector and, if desirable, to provide amplification within the host. Suitable prokaryotic hosts for transformation include *E. coli*, *Bacillus subtilis*, *Salmonella typhimurium* and various species within the genera *Pseudomonas*, *Streptomyces*, and *Staphylococcus*, although others may, also be employed as a matter of choice.

Bacterial vectors may be, for example, bacteriophage-, plasmid- or cosmid-based. These vectors can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids typically containing elements of the well known cloning vector pBR322 (ATCC 37017). Such commercial vectors include, for example, GEM 1 (Promega Biotec, Madison, WI, USA), pBs, phagescript, PsiX174, pBluescript SK, pBs KS, pNH8a, pNH16a, pNH18a, pNH46a (Stratagene); pTrc99A, pKK223-3, pKK233-3, pKK232-8, pDR540, and pRIT5 (Pharmacia).

These "backbone" sections are combined with an appropriate promoter and the structural sequence to be expressed. Bacterial promoters include lac, T3, T7, lambda P_R or P_L, trp, and ara.

Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter is derepressed/induced by appropriate means (e.g., temperature shift or chemical induction) and cells are cultured for an additional period. Cells are typically harvested by

centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

In bacterial systems, a number of expression vectors may be advantageously selected depending upon the use intended for the protein being expressed. For example, when a large quantity of such a protein is to be produced, for the generation of antibodies or to screen peptide libraries, for example, vectors which direct the expression of high levels of fusion protein products that are readily purified may be desirable. Such vectors include, but are not limited, to the *E. coli* expression vector pUR278 (Ruther et al., 1983, *EMBO J.* 2:1791), in which the coding sequence may be ligated into the vector in frame with the *lac Z* coding region so that a fusion protein is produced; pIN vectors (Inouye et al., 1985, *Nucleic Acids Res.* 13:3101-3109; Van Heeke et al., 1989, *J. Biol. Chem.* 264:5503-5509); pET vectors, Studier et al., *Methods in Enzymology* 185: 60-89 (Academic Press 1990); and the like.

Moreover, pGEX vectors may be used to express foreign polypeptides as fusion proteins with glutathione S-transferase (GST). In general, such fusion proteins are soluble and easily can be purified from lysed cells by adsorption to glutathione-agarose beads followed by elution in the presence of free glutathione. The pGEX vectors are designed to include thrombin or factor Xa protease cleavage sites so that the cloned target gene protein can be released from the GST moiety.

In a one embodiment, full length cDNA sequences are appended with in-frame *Bam*HI sites at the amino terminus and *Eco*RI sites at the carboxyl terminus using standard PCR methodologies (Innis et al., 1990, *supra*) and ligated into the pGEX-2TK vector (Pharmacia, Uppsala, Sweden). The resulting cDNA construct contains a kinase recognition site at the amino terminus for radioactive labeling and glutathione S-transferase sequences at the carboxyl terminus for affinity purification (Nilsson, et al. 1985, *EMBO J.* 4: 1075; Zabeau and Stanley, 1982, *EMBO J.* 1:1217).

Eukaryotic Expression

Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts,

described by Gluzman, *Cell* 23:175 (1981), and other cell lines capable of expressing a compatible vector, for example, the C127, 3T3, CHO, HeLa and BHK cell lines. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking nontranscribed sequences. DNA sequences derived from the SV40 viral genome, for example, SV40 origin, early promoter, enhancer, splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements.

Mammalian promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Exemplary mammalian vectors include pWLneo, pSV2cat, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, and pSVL (Pharmacia). Selectable markers include CAT (chloramphenicol transferase).

In mammalian host cells, a number of viral-based expression systems may be utilized. In cases where an adenovirus is used as an expression vector, the coding sequence of interest may be ligated to an adenovirus transcription/translation control complex, e.g., the late promoter and tripartite leader sequence. This chimeric gene may then be inserted in the adenovirus genome by *in vitro* or *in vivo* recombination. Insertion in a non-essential region of the viral genome (e.g., region E1 or E3) will result in a recombinant virus that is viable and capable of expressing a target protein in infected hosts. (E.g., See Logan et al., 1984, *Proc. Natl. Acad. Sci. USA* 81:3655-3659).

In one embodiment, cDNA sequences encoding the full-length open reading frames are ligated into pCMVB replacing the β -galactosidase gene such that cDNA expression is driven by the CMV promoter (Alam, 1990, *Anal. Biochem.* 188: 245-254; MacGregor et al., 1989, *Nucl. Acids Res.* 17: 2365; Norton et al. 1985, *Mol. Cell. Biol.* 5: 281).

In addition, a host cell strain may be chosen which modulates the expression of the inserted sequences, or modifies and processes the gene product in the specific fashion desired. Such modifications (e.g., glycosylation) and processing (e.g.,

cleavage) of protein products may be important for the function of the protein. Different host cells have characteristic and specific mechanisms for the post-translational processing and modification of proteins.

5 Appropriate cell lines or host systems can be chosen to ensure the correct modification and processing of the foreign protein expressed. To this end, eukaryotic host cells which possess the cellular machinery for proper processing of the primary transcript, glycosylation, and phosphorylation of the gene
10 product may be used. Such mammalian host cells include but are not limited to CHO, VERO, BHK, HeLa, COS, MDCK, 293, 3T3, WI38, etc.

15 For long-term, high-yield production of recombinant proteins in eukaryotic cells, stable expression is preferred. Rather than using expression vectors which contain viral origins of replication, host cells can be transformed with DNA controlled by appropriate expression control elements (e.g., promoter, enhancer, sequences, transcription terminators, polyadenylation sites, etc.), and a selectable marker.

20 Following the introduction of the foreign DNA, engineered cells may be allowed to grow for 1-2 days in an enriched media, and then are switched to a selective media. The selectable marker in the recombinant plasmid confers resistance to the selection and allows cells to stably integrate the plasmid into their
25 chromosomes and grow to form foci which in turn can be cloned and expanded into cell lines. This method may advantageously be used to engineer cell lines which express the target protein. Such engineered cell lines may be particularly useful in screening and evaluation of compounds that affect the endogenous activity of the
30 protein.

35 A number of selection systems may be used, including but not limited to the herpes simplex virus thymidine kinase (Wigler, et al., *Cell* 11:223 (1977)), hypoxanthine-guanine phosphoribosyltransferase (Szybalska et al., *Proc. Natl. Acad. Sci. USA* 48:2026 (1962)), and adenine phosphoribosyltransferase (Lowy, et al., *Cell* 22:817 (1980)) genes can be employed in tk⁻, hgprt⁻ or aprt⁻ cells, respectively. Also, antimetabolite resistance can be used as the basis of selection for dhfr, which

confers resistance to methotrexate (Wigler, et al., *Proc. Natl. Acad. Sci. USA* 77:3567 (1980)); O'Hare, et al., 1981, *Proc. Natl. Acad. Sci. USA* 78:1527); gpt, which confers resistance to mycophenolic acid (Mulligan et al., *Proc. Natl. Acad. Sci. USA* 78:2072 (1981)); neo, which confers resistance to the aminoglycoside G-418 (Colberre-Garapin, et al., 1981, *J. Mol. Biol.* 150:1); and hydro, which confers resistance to hygromycin (Santerre, et al., 1984, *Gene* 30:147) genes.

An alternative fusion protein system allows for the ready purification of non-denatured fusion proteins expressed in human cell lines (Janknecht, et al., *Proc. Natl. Acad. Sci. USA* 88:8972-8976 (1991)). In this system, the gene of interest is subcloned into a vaccinia-based plasmid such that the gene's open reading frame is translationally fused to an amino-terminal tag consisting of six histidine residues. Extracts from cells infected with recombinant vaccinia virus are loaded onto Ni²⁺ nitriloacetic acid-agarose columns and histidine-tagged proteins are selectively eluted with imidazole-containing buffers.

In an insect system, *Autographa californica* nuclear polyhedrosis virus (AcNPV) is used as a vector to express foreign genes. The virus grows in *Spodoptera frugiperda* cells. The target coding sequence may be cloned individually into non-essential regions (for example the polyhedrin gene) of the virus and placed under control of an AcNPV promoter (for example the polyhedrin promoter). Successful insertion of a target gene coding sequence will result in inactivation of the polyhedrin gene and production of non-occluded recombinant virus (i.e., virus lacking the proteinaceous coat coded for by the polyhedrin gene). These recombinant viruses are then used to infect *Spodoptera frugiperda* cells in which the inserted gene is expressed. (E.g., see Smith et al., 1983, *J. Virol.* 46: 584; Smith, U.S. Patent No. 4,215,051).

While the present proteins can be expressed in recombinant systems, as described above, cell-free translation systems can also be employed to produce such proteins using RNAs derived from the DNA constructs of the present invention.

Purification of Recombinant Proteins

Recombinant proteins produced may be isolated by host cell lysis. This may be followed by one or more salting-out, aqueous ion exchange or size exclusion chromatography steps. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps. Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents, like lysozyme and chelators.

If inclusion bodies are formed in bacterial systems, they may be extracted from cell pellets using, for example, detergents, reducing agents, salts, urea, guanidinium chloride and extremes of pH (e.g. <4 or >10). If denaturation occurs, protein refolding steps (e.g., dialysis) can be used, as necessary, in completing configuration of the mature protein. If disulfide bridges are present in the native protein, they may be reoxidized using known methods.

By way of specific non-limiting example, the recombinant bacterial cells, for example *E. coli*, are grown in any of a number of suitable media, for example LB, and the expression of the recombinant protein induced by adding IPTG (e.g., lac operator-promoter) to the media or switching incubation to a higher temperature (e.g., λ cI⁸⁵⁷). After culturing the bacteria for a further period of between 2 and 24 hours, the cells are collected by centrifugation and washed to remove residual media. The bacterial cells are then lysed, for example, by disruption in a cell homogenizer and centrifuged to separate the cell membranes from the soluble cell components. If the protein aggregates into inclusion bodies, this centrifugation can be performed under conditions whereby the dense inclusion bodies are selectively enriched by incorporation of sugars such as sucrose into the buffer and centrifugation at a selective speed. The inclusion bodies can then be washed in any of several solutions to remove some of the contaminating host proteins, then solubilized in solutions containing high concentrations of urea (e.g. 8M) or chaotropic agents such as guanidinium hydrochloride in the presence of reducing agents such as β -mercaptoethanol or DTT (dithiothreitol).

At this stage it may be advantageous to incubate the protein for several hours under conditions suitable for the protein to undergo a refolding process into a conformation which more closely resembles that of the native protein. Such conditions generally include low protein concentrations less than 500 µg/ml), low levels of reducing agent, concentrations of urea less than 2 M and often the presence of reagents such as a mixture of reduced and oxidized glutathione which facilitate the interchange of disulphide bonds within the protein molecule. The refolding process can be monitored, for example, by SDS-PAGE or with antibodies which are specific for the native molecule. Following refolding, the protein can then be purified further and separated from the refolding mixture by chromatography on any of several supports including ion exchange resins, gel permeation resins or on a variety of affinity columns.

Labeling Proteins

When used as a component in assay systems such as those described, below, the target protein may be labeled, either directly or indirectly, to facilitate detection of the present res-like molecules either *in vitro* or *in vivo*. Any of a variety of suitable labeling systems may be used including but not limited to radioisotopes such as ¹²⁵I; enzyme labeling systems that generate a detectable colorimetric signal or light when exposed to substrate; and fluorescent labels.

Where recombinant DNA technology is used for protein production then it may be advantageous to engineer fusion proteins that can facilitate labeling, immobilization and/or detection. These fusion proteins may, for example, add amino acids which facilitate further chemical modification. They also may add a functional moiety, such as an enzyme, which directly facilitates detection.

TRANSGENIC ANIMALS

The invention further contemplates animal models for studying the function of the present molecules and for overproducing the protein products. The disclosed DNA sequences may be used in conjunction with techniques for producing transgenic animals that are well known to those of skill in the art.

To prepare transgenic animals, target gene sequences may for example be introduced into, and overexpressed in, the genome of the animal of interest, or, if endogenous target gene sequences are present, they may either be overexpressed or, alternatively, be disrupted in order to underexpress or inactivate target gene expression, such as described for the disruption of apoE in mice (Plum et al., Cell 71: 343-353 (1992)).

In order to overexpress a target gene sequence, the coding portion of the target gene sequence may be ligated to a regulatory sequence which is capable of driving gene expression in the animal and cell type of interest. Such regulatory regions will be well known to those of skill in the art, and may be utilized in the absence of undue experimentation.

For underexpression of an endogenous target gene sequence, such a sequence may be isolated and engineered such that when reintroduced into the genome of the animal of interest, the endogenous target gene alleles will be inactivated. Preferably, the engineered target gene sequence is introduced via gene targeting such that the endogenous target sequence is disrupted upon integration of the engineered target gene sequence into the animal's genome. Animals of any species, including, but not limited to, mice, rats, rabbits, guinea pigs, pigs, micro-pigs, goats, and non-human primates, e.g., baboons, monkeys, and chimpanzees may be used to generate cardiovascular disease animal models. Goats, cows and sheep are particularly preferred for producing protein *in vivo*.

Any technique known in the art may be used to introduce a target gene transgene into animals to produce the founder lines of transgenic animals. Such techniques include, but are not limited to pronuclear microinjection (Hoppe et al., U.S. Pat. No. 4,873,191 (1989)); retrovirus mediated gene transfer into germ lines (Van der Putten et al., Proc. Natl. Acad. Sci., USA 82:6148-

6152 (1985)); gene targeting in embryonic stem cells (Thompson et al., *Cell* 56:313-321 (1989)); electroporation of embryos (Lo, *Mol. Cell. Biol.* 3:1803-1814 (1983)); and sperm-mediated gene transfer (Lavitrano et al., *Cell* 57:717-723 (1989)); etc. For a review of such techniques, see Gordon, *Transgenic Animals*, *Intl. Rev. Cytol.* 115:171-229 (1989).

The present invention provides for transgenic animals that carry the transgene in all their cells, as well as animals which carry the transgene in some, but not all their cells, i.e., mosaic animals. The transgene may be integrated as a single transgene or in concatamers, e.g., head-to-head tandems or head-to-tail tandems. The transgene may also be selectively introduced into and activated in a particular cell type by following, for example, the teaching of Lasko et al. (Lasko et al., *Proc. Natl. Acad. Sci. USA* 89:3232-6236 (1992)). The regulatory sequences required for such a cell-type specific activation will depend upon the particular cell type of interest, and will be apparent to those of skill in the art. When it is desired that the target gene be integrated into the chromosomal site of the endogenous target gene, gene targeting is preferred. Briefly, when such a technique is to be utilized, vectors containing some nucleotide sequences homologous to the endogenous target gene of interest are designed for the purpose of integrating, via homologous recombination with chromosomal sequences, into and disrupting the function of the nucleotide sequence of the endogenous target gene.

The transgene may also be selectively introduced into a particular cell type, thus inactivating the endogenous gene of interest in only that cell type, by following, for example, the teaching of Gu et al. *Science* 265: 103-106 (1994)). The regulatory sequences required for such a cell-type specific inactivation will depend upon the particular cell type of interest, and will be apparent to those of skill in the art.

Once transgenic animals have been generated, the expression of the recombinant target gene and protein may be assayed utilizing standard techniques. Initial screening may be accomplished by Southern blot analysis or PCR techniques to analyze animal tissues to assay whether integration of the transgene has taken place. The level of mRNA expression of the

transgene in the tissues of the transgenic animals may also be assessed using techniques which include but are not limited to Northern blot analysis of tissue samples obtained from the animal, in situ hybridization analysis, and RT-PCR. Samples of target gene-expressing tissue, may also be evaluated immunocytochemically using antibodies specific for the target gene transgene gene product of interest.

The transgenic animals that express target gene mRNA or target gene transgene peptide (detected immunocytochemically, using antibodies directed against the target gene product's epitopes) at easily detectable levels should then be further evaluated to identify those animals which display characteristic increased susceptibility to carcinogenesis. Additionally, specific cell types within the transgenic animals may be analyzed and assayed *in vitro* for cellular phenotypes characteristic of mutant phenotype.

Once target gene transgenic founder animals are produced, they may be bred, inbred, outbred, or crossbred to produce colonies of the particular animal. Examples of such breeding strategies include but are not limited to: outbreeding of founder animals with more than one integration site in order to establish separate lines; inbreeding of separate lines in order to produce compound target gene transgenics that express the target gene transgene of interest at higher levels because of the effects of additive expression of each target gene transgene; crossing of heterozygous transgenic animals to produce animals homozygous for a given integration site in order both to augment expression and eliminate the possible need for screening of animals by DNA analysis; crossing of separate homozygous lines to produce compound heterozygous or homozygous lines; breeding animals to different inbred genetic backgrounds so as to examine effects of modifying alleles on expression of the target gene transgene and the possible development of carcinogenesis. One such approach is to cross the target gene transgenic founder animals with a wild type strain to produce an F₁ generation that exhibits increased susceptibility to carcinogenesis. The F₁ generation may then be inbred in order to develop a homozygous line, if it is found that homozygous target gene transgenic animals are viable.

Methods of generating "knockout" mice using homologous recombination in embryonic stem cells are well known in the art. Suitable methods are described, for example, in Mansour et al., *Nature*, 336:348 (1988); Zijlstra et al., *Nature*, 342:435 (1989) and 344:742 (1990); and Hasty et al., *Nature*, 350:243 (1991). This genomic DNA can be obtained by conventional methods using the cDNA sequence as a probe in a commercially-available genomic DNA library.

Briefly, a genomic fragment is cleaved with a restriction endonuclease and a heterologous cassette containing a neomycin-resistance gene is inserted at the cleavage site. A suitable cassette is the GTI-II neo cassette described by Lufkin et al., *Cell* 66:1105 (1991). The modified genomic fragment is cloned into a suitable targeting vector that is introduced into murine embryonic stem cells by electroporation. Cells that have undergone homologous recombination (and hence disruption of the gene) are selected by resistance to G418, and used to generate chimeric mice using well known methods. See Lufkin et al., *supra*. Traditional breeding methods then can be used to generate mice that are homozygous for the disrupted gene.

The phenotype of mice that are homozygous for the mutation then can be studied to provide insights into the role of the protein in, for example, carcinogenesis. These mice also can be used as models for developing new treatments for cancers. If this mutation is lethal in homozygous mice (for example during embryogenesis) heterozygous mice, which express only half the amount of the protein can also be studied.

GENE THERAPY APPLICATIONS

When mutations in the inventive protein, or in the elements controlling expression of that protein, are found to be associated with a malignant phenotype, control of cellular proliferation can be restored by gene therapy methods. For example, overexpression of the protein can be counteracted by concurrent expression of an antisense molecule that binds to and inhibits expression of the mRNA encoding the protein. Alternatively, overexpression can be inhibited in an analogous manner using a ribozyme that cleaves the mRNA. In another embodiment, where expression of a mutated

protein induces the malignant phenotype, concomitant expression of the non-mutated molecule via introduction of an exogenous gene may be used. Methods of using antisense and ribozyme technology to control gene expression, or of gene therapy methods for expression of an exogenous gene in this manner are well known in the art.

Each of these methods requires a system for introducing a vector into the cells containing the mutated gene. The vector encodes either an antisense or ribozyme transcript of the inventive protein. The construction of a suitable vector can be achieved by any of the methods well-known in the art for the insertion of exogenous DNA into a vector. See, e.g., Sambrook et al., *Molecular Cloning* (Cold Spring Harbor Press 2d ed. 1989), which is incorporated herein by reference. In addition, the prior art teaches various methods of introducing exogenous genes into cells *in vivo*. See Rosenberg et al., *Science* 242:1575-1578 (1988) and Wolff et al., *PNAS* 86:9011-9014 (1989), which are incorporated herein by reference. The routes of delivery include systemic administration and administration *in situ*. Well-known techniques include systemic administration with cationic liposomes, and administration *in situ* with viral vectors. Any one of the gene delivery methodologies described in the prior art is suitable for the introduction of a recombinant vector containing an inventive gene according to the invention into a MTX-resistant, transport-deficient cancer cell. A listing of present-day vectors suitable for the purpose of this invention is set forth in Hodgson, *Bio/Technology* 13: 222 (1995), which is incorporated by reference.

For example, liposome-mediated gene transfer is a suitable method for the introduction of a recombinant vector containing an inventive gene according to the invention into a MTX-resistant, transport-deficient cancer cell. The use of a cationic liposome, such as DC-Chol/DOPE liposome, has been widely documented as an appropriate vehicle to deliver DNA to a wide range of tissues through intravenous injection of DNA/cationic liposome complexes. See Caplen et al., *Nature Med.* 1:39-46 (1995) and Zhu et al., *Science* 261:209-211 (1993), which are herein incorporated by reference. Liposomes transfer genes to the target cells by fusing with the plasma membrane. The entry process is relatively efficient, but once inside the cell, the liposome-DNA complex has

no inherent mechanism to deliver the DNA to the nucleus. As such, the most of the lipid and DNA gets shunted to cytoplasmic waste systems and destroyed. The obvious advantage of liposomes as a gene therapy vector is that liposomes contain no proteins, which thus minimizes the potential of host immune responses.

As another example, viral vector-mediated gene transfer is also a suitable method for the introduction of the vector into a target cell. Appropriate viral vectors include adenovirus vectors and adeno-associated virus vectors, retrovirus vectors and herpesvirus vectors.

Adenoviruses are linear, double stranded DNA viruses complexed with core proteins and surrounded by capsid proteins. The common serotypes 2 and 5, which are not associated with any human malignancies, are typically the base vectors. By deleting parts of the virus genome and inserting the desired gene under the control of a constitutive viral promoter, the virus becomes a replication deficient vector capable of transferring the exogenous DNA to differentiated, non-proliferating cells. To enter cells, the adenovirus fibre interacts with specific receptors on the cell surface, and the adenovirus surface proteins interact with the cell surface integrins. The virus penton-cell integrin interaction provides the signal that brings the exogenous gene-containing virus into a cytoplasmic endosome. The adenovirus breaks out of the endosome and moves to the nucleus, the viral capsid falls apart, and the exogenous DNA enters the cell nucleus where it functions, in an epichromosomal fashion, to express the exogenous gene. Detailed discussions of the use of adenoviral vectors for gene therapy can be found in Berkner, *Biotechniques* 6:616-629 (1988) and Trapnell, *Advanced Drug Delivery Rev.* 12:185-199 (1993), which are herein incorporated by reference. Adenovirus-derived vectors, particularly non-replicative adenovirus vectors, are characterized by their ability to accommodate exogenous DNA of 7.5 kB, relative stability, wide host range, low pathogenicity in man, and high titers (10^4 to 10^5 plaque forming units per cell). See Stratford-Perricaudet et al., *PNAS* 89:2581 (1992).

Adeno-associated virus (AAV) vectors also can be used for the present invention. AAV is a linear single-stranded DNA parvovirus

that is endogenous to many mammalian species. AAV has a broad host range despite the limitation that AAV is a defective parvovirus which is dependent totally on either adenovirus or herpesvirus for its reproduction *in vivo*. The use of AAV as a vector for the introduction into target cells of exogenous DNA is well-known in the art. See, e.g., Lebkowski et al., *Mole. & Cell. Biol.* 8:3988 (1988), which is incorporated herein by reference. In these vectors, the capsid gene of AAV is replaced by a desired DNA fragment, and transcomplementation of the deleted capsid function is used to create a recombinant virus stock. Upon infection the recombinant virus uncoats in the nucleus and integrates into the host genome.

Another suitable virus-based gene delivery mechanism is retroviral vector-mediated gene transfer. In general, retroviral vectors are well-known in the art. See Breakfield et al., *Mole. Neuro. Biol.* 1:339 (1987) and Shih et al., in *Vaccines* 85: 177 (Cold Spring Harbor Press 1985). A variety of retroviral vectors and retroviral vector-producing cell lines can be used for the present invention. Appropriate retroviral vectors include Moloney Murine Leukemia Virus, spleen necrosis virus, and vectors derived from retroviruses such as Rous Sarcoma Virus, Harvey Sarcoma Virus, avian leukosis virus, human immunodeficiency virus, myeloproliferative sarcoma virus, and mammary tumor virus. These vectors include replication-competent and replication-defective retroviral vectors. In addition, amphotropic and xenotropic retroviral vectors can be used. In carrying out the invention, retroviral vectors can be introduced to a tumor directly or in the form of free retroviral vector producing-cell lines. Suitable producer cells include fibroblasts, neurons, glial cells, keratinocytes, hepatocytes, connective tissue cells, ependymal cells, chromaffin cells. See Wolff et al., *PNAS* 84:3344 (1989).

Retroviral vectors generally are constructed such that the majority of its structural genes are deleted or replaced by exogenous DNA of interest, and such that the likelihood is reduced that viral proteins will be expressed. See Bender et al., *J. Virol.* 61:1639 (1987) and Armento et al., *J. Virol.* 61:1647 (1987), which are herein incorporated by reference. To facilitate expression of the antisense or ribozyme molecule, of the inventive

protein, a retroviral vector employed in the present invention must integrate into the genome of the host cell genome, an event which occurs only in mitotically active cells. The necessity for host cell replication effectively limits retroviral gene
5 expression to tumor cells, which are highly replicative, and to a few normal tissues. The normal tissue cells theoretically most likely to be transduced by a retroviral vector, therefore, are the endothelial cells that line the blood vessels that supply blood to the tumor. In addition, it is also possible that a retroviral
10 vector would integrate into white blood cells both in the tumor or in the blood circulating through the tumor.

The spread of retroviral vector to normal tissues, however, is limited. The local administration to a tumor of a retroviral vector or retroviral vector producing cells will restrict vector
15 propagation to the local region of the tumor, minimizing transduction, integration, expression and subsequent cytotoxic effect on surrounding cells that are mitotically active.

Both replicatively deficient and replicatively competent retroviral vectors can be used in the invention, subject to their
20 respective advantages and disadvantages. For instance, for tumors that have spread regionally, such as lung cancers, the direct injection of cell lines that produce replication-deficient vectors may not deliver the vector to a large enough area to completely eradicate the tumor, since the vector will be released only from
25 the original producer cells and their progeny, and diffusion is limited. Similar constraints apply to the application of replication deficient vectors to tumors that grow slowly, such as human breast cancers which typically have doubling times of 30 days versus the 24 hours common among human gliomas. The much
30 shortened survival-time of the producer cells, probably no more than 7-14 days in the absence of immunosuppression, limits to only a portion of their replicative cycle the exposure of the tumor cells to the retroviral vector.

The use of replication-defective retroviruses for treating
35 tumors requires producer cells and is limited because each replication-defective retrovirus particle can enter only a single cell and cannot productively infect others thereafter. Because these replication-defective retroviruses cannot spread to other tumor cells, they would be unable to completely penetrate a deep,

multilayered tumor *in vivo*. See Markert et al., *Neurosurg.* 77: 590 (1992). The injection of replication-competent retroviral vector particles or a cell line that produces a replication-competent retroviral vector virus may prove to be a more effective therapeutic because a replication competent retroviral vector will establish a productive infection that will transduce cells as long as it persists. Moreover, replicatively competent retroviral vectors may follow the tumor as it metastasizes, carried along and propagated by transduced tumor cells. The risks for complications are greater, with replicatively competent vectors, however. Such vectors may pose a greater risk than replicatively deficient vectors of transducing normal tissues, for instance. The risks of undesired vector propagation for each type of cancer and affected body area can be weighed against the advantages in the situation of replicatively competent versus replicatively deficient retroviral vector to determine an optimum treatment.

Both amphotropic and xenotropic retroviral vectors may be used in the invention. Amphotropic viruses have a very broad host range that includes most or all mammalian cells, as is well known to the art. Xenotropic viruses can infect all mammalian cell cells except mouse cells. Thus, amphotropic and xenotropic retroviruses from many species, including cows, sheep, pigs, dogs, cats, rats, and mice, *inter alia* can be used to provide retroviral vectors in accordance with the invention, provided the vectors can transfer genes into proliferating human cells *in vivo*.

Clinical trials employing retroviral vector therapy treatment of cancer have been approved in the United States. See Culver, *Clin. Chem.* 40: 510 (1994). Retroviral vector-containing cells have been implanted into brain tumors growing in human patients. See Oldfield et al., *Hum. Gene Ther.* 4: 39 (1993). These retroviral vectors carried the HSV-1 thymidine kinase (HSV-tk) gene into the surrounding brain tumor cells, which conferred sensitivity of the tumor cells to the antiviral drug ganciclovir. Some of the limitations of current retroviral based cancer therapy, as described by Oldfield are: (1) the low titer of virus produced, (2) virus spread is limited to the region surrounding the producer cell implant, (3) possible immune response to the producer cell line, (4) possible insertional mutagenesis and

transformation of retroviral infected cells, (5) only a single treatment regimen of pro-drug, ganciclovir, is possible because the "suicide" product kills retrovirally infected cells and producer cells and (b) the bystander effect is limited to cells in direct contact with retrovirally transformed cells. See Bi et al., *Human Gene Therapy* 4: 725 (1993).

Yet another suitable virus-based gene delivery mechanism is herpesvirus vector-mediated gene transfer. While much less is known about the use of herpesvirus vectors, replication-competent HSV-1 viral vectors have been described in the context of antitumor therapy. See Martuza et al., *Science* 252: 854 (1991), which is incorporated herein by reference.

DIAGNOSTIC METHODS

The present invention also contemplates, for certain molecules described below, methods for diagnosis of human disease. In particular, patients can be screened for the occurrence of cancers, or likelihood of occurrence of cancers, associated with mutations in the encoded protein. DNA from tumor tissue obtained from patients suffering from cancer can be isolated and the gene encoding the protein can be sequenced. By examining a number of patients in this manner, mutations in the gene that are associated with a malignant cellular phenotype can be identified. In addition, correlation of the nature of the observed mutations with subsequent observed clinical outcomes allows development of prognostic model for the predicted outcome in a particular patient.

Screening for mutations conveniently can be carried out at the DNA level by use of PCR, although the skilled artisan will be aware that many other well known methods are available for the screening. PCR primers can be selected that flank known mutation sites, and the PCR products can be sequenced to detect the occurrence of the mutation. Alternatively, the 3' residue of one PCR primer can be selected to be a match only for the residue found in the unmutated gene. If the gene is mutated, there will be a mismatch at the 3' end of the primer, and primer extension cannot occur, and no PCR product will be obtained. Alternatively, primer mixtures can be used where the 3' residue of one primer is

any nucleotide other than the nonmutated residue. Observation of a PCR product then indicates that a mutation has occurred. Other methods of using, for example, oligonucleotide probes to screen for mutations are described, for example, in U.S. Patent No. 4,871,838, which is herein incorporated by reference in its entirety.

Alternatively, antibodies can be generated that selectively bind either mutated or non-mutated protein. The antibodies then can be used to screen tissue samples for occurrence of mutations in a manner analogous to the DNA-based methods described *supra*.

The diagnostic methods described above can be used not only for diagnosis and for prognosis of existing disease, but may also be used to predict the likelihood of the future occurrence of disease. For example, clinically healthy patients can be screened for mutations in the inventive molecule that correlate with later disease onset. Such mutations may be observed in the heterozygous state in healthy individuals. In such cases a single mutation event can effectively disable proper functioning of the gene and induce a transformed or malignant phenotype. This screening also may be carried out prenatally or neonatally.

DNA molecules according to the invention also are well suited for use in so-called "gene chip" diagnostic applications. Such applications have been developed by, *inter alia*, Synteni and Affymetrix. Briefly, all or part of the DNA molecules of the invention can be used either as a probe to screen a polynucleotide array on a "gene chip," or they may be immobilized on the chip itself and used to identify other polynucleotides via hybridization to the surface of the chip. In this manner, for example, related genes can be identified, or expression patterns of the gene in various tissues can be simultaneously studied. Such gene chips have particular application for diagnosis of disease, or in forensic analysis to detect the presence or absence of an analyte. Suitable chip technology is described for example, in Wodicka et al., *Nature Biotechnology*, 15:1359 (1997) which is hereby incorporated by reference in its entirety, and references cited therein.

PROTEIN-PROTEIN INTERACTIONS

Due to their similarity to certain known proteins, it is anticipated that some of the inventive protein molecules will interact with another class of cellular proteins. This is particularly true of those molecule containing leucine zipper motifs.

Any method suitable for detecting protein-protein interactions can be employed for identifying interacting targets. Among the traditional methods which can be employed are co-immunoprecipitation, crosslinking and co-purification through gradients or chromatographic columns. Utilizing procedures such as these allows for the identification of GAP gene products. Once identified, a GAP protein can be used, in conjunction with standard techniques, to identify its corresponding pathway gene. For example, at least a portion of the amino acid sequence of the pathway gene product can be ascertained using techniques well known to those of skill in the art, such as via the Edman degradation technique (see, e.g., Creighton, 1983, PROTEINS: STRUCTURES AND MOLECULAR PRINCIPLES, W.H. Freeman & Co., N.Y., pp.34-49). The amino acid sequence obtained can be used as a guide for the generation of oligonucleotide mixtures that can be used to screen for pathway gene sequences. Screening can be accomplished, for example, by standard hybridization or PCR techniques. Techniques for the generation of oligonucleotide mixtures and for screening are well-known. (See e.g., Ausubel, *supra*, and PCR PROTOCOLS: A GUIDE TO METHODS AND APPLICATIONS, 1990, Innis et al., eds. Academic Press, Inc., New York).

Additionally, methods can be employed which result in the simultaneous identification of interacting target genes. One method which detects protein interactions *in vivo*, the two-hybrid system, is described in detail for illustration purposes only and not by way of limitation. One version of this system has been described (Chien et al., *Proc. Natl. Acad. Sci. USA*, 88: 9578-9582 (1991)) and is commercially available from Clontech (Palo Alto, CA).

Briefly, utilizing such a system, plasmids are constructed that encode two hybrid proteins: one consists of the DNA-binding domain of a transcription activator protein fused to a known protein, in this case an inventive protein, and the other contains

the activator protein's activation domain fused to an unknown protein (a putative GAP, for instance) that is encoded by a cDNA which has been recombined into this plasmid as part of a cDNA library. The plasmids are transformed into a strain of the yeast
5 *Saccharomyces cerevisiae* that contains a reporter gene (e.g., *lacZ*) whose regulatory region contains the transcription activator's binding sites. Either hybrid protein alone cannot activate transcription of the reporter gene, the DNA-binding domain hybrid cannot because it does not provide activation
10 function, and the activation domain hybrid cannot because it cannot localize to the activator's binding sites. Interaction of the two hybrid proteins reconstitutes the functional activator protein and results in expression of the reporter gene, which is detected by an assay for the reporter gene product.

15 The two-hybrid system or related methodology can be used to screen activation domain libraries for proteins that interact with a known "bait" gene product. By way of example, and not by way of limitation, gene products known to be involved in TH cell subpopulation-related disorders and/or differentiation,
20 maintenance, and/or effector function of the subpopulations can be used as the bait gene products. Total genomic or cDNA sequences are fused to the DNA encoding on activation domain. This library and a plasmid encoding a hybrid of the bait gene product fused to the DNA-binding domain are cotransformed into a yeast reporter
25 strain, and the resulting transformants are screened for those that express the reporter gene. For example, and not by way of limitation, the bait gene can be cloned into a vector such that it is translationally fused to the DNA encoding the DNA-binding domain of the GAL4 protein. These colonies are purified and the
30 library plasmids responsible for reporter gene expression are isolated. DNA sequencing is then used to identify the proteins encoded by the library plasmids.

The present invention, thus generally described, will be
35 understood more readily by reference to the following examples, which are provided by way of illustration and are not intended to be limiting of the present invention.

The examples below are provided to illustrate the subject invention. These examples are provided by way of illustration and are not included for the purpose of limiting the invention.

EXAMPLES

5 EXAMPLE I: cDNA Library Construction

cDNA library plates and clones originated from five cDNA libraries that were constructed by directional cloning. These are available through the Resource Center (<http://www.rzpd.de>) of the
10 German Genome Project. In particular, the hfbr2 (human fetal brain; RZPD number DKFZp564) and hfkd2 (human fetal kidney; DKFZp566) libraries were generated using the Smart kit (Clontech), except that PCR was carried out with primers that contained uracil residues to permit directional cloning without
15 restriction digestion and ligation, and were complementary with the pAMP1 (LifeTechnologies) cloning sites for directional cloning. The htes3 (human testes; DKFZp434), hute1 (human uterus; DKFZp586) and hmcfl (human mammary carcinoma; DKFZp727) libraries are conventional (Gubler, U., Hoffman, B.J., (1983), A simple and
20 very efficient method for generating cDNA libraries. Gene 25, 263-269), size-selected cDNA libraries. They are cloned into pSPORT1 (LifeTechnologies) via a NotI site which is introduced during reverse transcription downstream of the oligo dT primer and a SalI site that is introduced by the ligation of a adapters.
25 The human mammary carcinoma library was constructed from MCF7 cells.

In a similar fashion, the hamy2 (human amygdala nucleus (inside the brain); RZPD number DKFZp761) and hmel2 (human melanoma; RZPD number DKFZp762) libraries have been generated
30 using conventional approaches, employing a NotI -dT V primer for first strand synthesis (GAGCGGCCGC(T)19V). After second strand synthesis, SalI adapters were ligated to the blunted cDNA. Then the cDNA was cut with NotI to generate SalI-NotI compatible ends at the 5' and 3' ends of the cDNA, respectively, to allow
35 directional cloning. The cDNAs were then size selected on agarose gels in two dimensions and cloned into the pSPORT1 plasmid vector which had been pre-cut with SalI and NotI (LifeTechnologies). The

DNA was transformed into the DH10B bacterial strain and single colonies were picked into 384well microtiter plates from the non-amplified library. The human melanoma library was constructed from MeWo cells, published by Kern, M.A., Helmbach, H., Artuc, M., Karmann, D., Jurgovsky, K. and Schadendorf, D. (1997) Human melanoma cell lines selected in vitro displaying various levels of drug resistance against cisplatin, fotemustine, vindesine or etoposide: modulation of proto-oncogene expression. Anticancer Res. 17, 4359-4370.

10 The cDNA sequences of this application were first identified among the sequences comprising various libraries. Technology has advanced considerably since the first cDNA libraries were made. Many small variations in both chemicals and machinery have been instituted over time, and these have improved both the efficiency
15 and safety of the process. Although the cDNAs could be obtained using an older procedure, the procedure presented in this application is exemplary of one currently being used by persons skilled in the art. For the purpose of providing an exemplary method, the mRNA isolation and cDNA library construction
20 described here is for the MCF-7 library (DKFZp727) from which the clones named DKFZphmcfl_xxyyxx were obtained.

The human cell line MCF-7 was grown in DMEM supplemented with 10% fetal calf serum until confluency. 3×10^5 cells were harvested with a cell scraper in PBS. Cells were lysed in buffer
25 containing 0.5 % NP-40 to leave the nuclei intact. The debris was pelleted by centrifugation at $15\,000 \times g$ for 10 minutes at 4 degrees Celsius. Proteins in the supernatant were degraded in presence of SDS and Proteinase K (30 minutes at 56 degrees Celsius). Precipitation of proteins was done in a
30 Phenol/Chloroform extraction, RNA was precipitated from the aqueous phase with Na-acetate and Ethanol. Polyadenylated messages were isolated using Qiagen Oligotex (QIAGEN, Hilden Germany).

First strand cDNA synthesis was accomplished using an oligo
35 (dT) primer which also contained an NotI restriction site. Second strand synthesis was performed using a combination of DNA polymerase I, E. coli ligase and RNase H, followed by the

addition of a SalI adaptor to the blunt ended cDNA. The SalI adapted, double-stranded cDNA was then digested with NotI restriction enzyme, and fractionated by size on an agarose gel. DNA of the appropriate size was cut from the gel and cast into a second gel in a 90° angle. After electrophoresis in the second dimension, cDNA of the appropriate size was cut from the gel. The agarose block was broken down with help of gelase. The cDNA was purified with help of two phenol extractions and an ethanol precipitation. The cDNA was ligated into SalI/NotI pre-digested pSport1 vector (LifeTechnologies) and transformed into DH10B bacteria.

The libraries were arrayed into 384-well microtiter plates and spotted on high density nylon membranes for hybridization analysis. All libraries have been arrayed into 384well microtiter plates and spotted on high density nylon membranes for hybridization analysis. The hamy2 Library consists of 121 384well plates comprising 46464 clones. The hmel2 library consists of 72 384well plates comprising 27648 clones. Filters and clones are available through the Resource Center of German Genome Project (<http://www.RZPD.de>). Whole library plates were distributed to the sequencing partners of the consortium for systematic sequencing.

25 EXAMPLE II: Sequencing of cDNA Clones

All clones in the 384-well microtiter plates were sequenced from the 5' end. Sequencing was done preferentially using dye terminator chemistry (ABD or Amersham) on ABI automated DNA sequencers (ABI 377, Applied Biosystems), one partner used EMBL prototype instruments (Arakis) mainly with dye primer chemistry.

The resulting expressed sequence tag (EST) sequences ("r1 ESTs" = sequenced from 5'-end) were analysed for:

a) the lack of identical matches with known genes.

For this, the EST-sequence was blasted against the cDNA consortiums own database and after that against public databases

and (with BLASTn and BLASTx against EMBL/EMBLNEW and assembled ESTs, please refer to EXAMPLE III: Bioinformatics analysis of full length cDNAs, for description and parameter settings). ESTs which were identical to known genes in more than 100 bp, with
5 less than 2 mismatches, were excluded from further analysis.

b) the presence of an open reading frame

Open reading frames (ORFs) were detected with a tool developed by Munich Information Center for Protein Sequences (MIPS) called ORF-map. ORF-map visualises potential start and
10 stop-codons. If an ORF without a stop codon was detected in a rl-EST, the sequence was processed further.

c) the presence of GC rich sequences

A script developed by MIPS computed the GC-content of the rl-sequence, which should be >40%. Writing similar scripts is
15 within the ordinary skill of one in bioinformatics.

d) the lack of repeat structures

Repeats such as Alu, Line or CA-repeats were detected by blasting (BLASTn and BLASTx, please refer to EXAMPLE III: Bioinformatics analysis of full length cDNAs, for description and
20 parameter settings) against a repeat-database compiled by MIPS. If a repeat was present within the rl-sequence, the sequence were not processed further.

Novel clones that met all criteria were identified to the sequencers, who then performed 3'-end sequencing of these clones.
25 The resulting 3' ESTs ("s1 ESTs" = sequenced from 3'-end) were checked for

a) the lack of matches with known genes in public databases, and sequences already generated by us.

This was done by blasting against EMBL/EMBLNEW and assembled
30 EST (BLASTn and BLASTx, please refer to EXAMPLE III: Bioinformatics analysis of full length cDNAs, for description and parameter settings).

b) the presence of polyadenylation signals.

Again only clones matching the selection criteria were chosen to be sequenced completely by the sequencers. Clones were selected after the following criteria:

5 A very good ORF had at least one BLASTx match to other proteins. A "good ORF" should extend to the 3' end and be longer than ~40 codons. If the ORF started in the r1 sequence, in front of the potential start codon, there should not exist too many competing start codons in frame with the ORF start codon and the
10 start should match the Kozak consensus ATG. If the EST sequence was too short to decide according to the potential ORF, and there were only a few or no start codons in the sequence the GC content of the Sequence should be greater than 40%. The r1 sequences needed not contain an polyA-tail at the 3' end. In addition, the
15 results of the blasting against the assembled human ESTs could help in questionable cases to decide whether to stop or to continue. A hit against these ESTs was an indication to go further.

Clones passing the above-described screening were sequenced
20 in full. Sequencing was done preferentially using dye terminator chemistry (ABD or Amersham) on ABI automated DNA sequencers (ABI 377; Applied Biosystems), one partner used EMBL prototype instruments (Arakis) mainly with dye primer chemistry. Primer walking (Strauss et al., 1986, Specific-primer-directed DNA
25 sequencing. Anal Biochem. 154, 353-360) was the preferred sequencing strategy because of the lower redundancy possible compared to random shotgun (Messing, J., Crea, R., Seeburg, H.P. (1981) A system for shotgun DNA sequencing. Nucleic Acids Res. 9, 32-39) methods. Walking primers were generally designed using
30 software (e.g. Haas, S., Vingron, M., Poustka, A., Wiemann, S. (1998) Primer design in large-scale sequencing. Nucleic Acids Res. 26, 3006-3012, Schwager, C., Wiemann, S., Ansorge, W. (1995) GeneSkipper: integrated software environment for DNA sequence assembly and alignment. HUGO Genome Digest 2, 8-9) that permitted
35 complete automation of this usually time consuming process and helped in the parallel processing of large numbers of clones.

EXAMPLE III: Bioinformatics analysis of full length cDNAs

Each sequence obtained was compared on nucleotide level in a stepwise manner to sequences in EMBL/EMBLNEW, EMBL-EST, EMBL-STS using the BLASTn algorithm. Basic Local Alignment Search Tool (BLAST, Altschul S. F. (1993) J Mol Evol 36:290-300; Altschul, S. F. et al (1990) J Mol Biol 215:403-10) is used to search for local sequence alignments. BLAST produces alignments of both nucleotide (BLASTn) and amino acid sequences (BLASTp or BLASTx) to determine sequence similarity. BLAST is especially useful in determining exact matches or in identifying homologs, because of the local nature of the alignments. While it is useful for matches which do not contain gaps, it is inappropriate for performing motif-style searching. The fundamental unit of BLAST algorithm output is the High-scoring Segment Pair (HSP).

An HSP consists of two sequence fragments of arbitrary but equal lengths whose alignment is locally maximal and for which the alignment BLAST approach is to look threshold or cut off score set by the user. BLAST looks for HSPs between a query sequence and a database sequence, to evaluate the statistical significance of any matches found, and to report only those matches which satisfy the user-selected threshold of significance. The parameter E establishes the statistically significant threshold for reporting database sequence matches. E is interpreted as the upper bound of the expected frequency of chance occurrence of an HSP (or set of HSPs) within the context of the entire database search. Any database sequence whose match satisfies E is reported in the program output. Parameter settings for the BLAST-operations (BLASTN 2.0a19MP-WashU) described were: EMBL-EMBLNEW: H=0 V=5 B=5 -filter seg; EMBL-EST: H=0 E=1e-10 B=500 V=500 -filter seg; EMBL-STS: H=0 V=5 B=5.

Search against EMBL/EMBLNEW was done to determine whether the cDNAs are already known, and also to find out whether the cDNAs are encoded by genomic sequences already sequenced and published/submitted to these databases.

Search against EMBL-EST was performed to get a first impression how abundant a particular cDNA would be and to get

information on tissue specificity (so-called "electronic Northern-Blot", e.g. some of the cDNAs derived of the testis library show only hits to ESTs also derived of testis libraries).

5 The cDNA-sequences were blasted against EMBL-STS to determine STS-sequence-match to the cDNA, thus providing a mapping information to the new cDNA.

10 The potential protein-sequences were generated automatically by a script searching for the longest open reading frame (ORF) in each of the three forward frames with a minimum length of 90 codons. Next, the automatically generated ORFs were translated into protein sequences. These protein sequences were searched against the non redundant protein data set of PIR/SwissProt/Trembel/Tremblnew (BLASTP 2.0a19MP-WashU, parameter setting: V=7 B=7 H=0 -filter seg). If the script generated more
15 than one ORF, one ORF was chosen manually by the annotater according to the degree of similarity to known proteins, the location of the ORF in the cDNA, the length, the amino acid composition and the content of Prosite-Motifs.

20 Additionally there was a BLASTx (BLASTX 2.0a19MP-WashU against non redundant protein database comprising PIR/SWISSPROT/TREMBL/TREMBLNEW; parameter-settings were: matrix/home/data/blast/matrix/aa/BL0SUM62 H=0 V=5 B=5 -filter seg) search to find potential frame shift in the complementary cds of the cDNAs and to identify unspliced or partly spliced
25 cDNAs. The protein sequence was then transferred to the PEDANT system, in order to generate additional information on the new proteins. PEDANT (Protein Extraction, Description, and ANalysis Tool, Frishman, D. & Mewes, H.-W. (1997) PEDANTic, genome analysis. Trends in Genetics, 13, 415-416) is a platform
30 developed at the Munich Information Center for Protein Sequences (MIPS, Munich, Germany), which incorporates practically all bioinformatics methods important for the functional and structural characterisation of protein sequences. Computational methods used by PEDANT are:

FASTA

Very sensitive protein sequence database searches with estimates of statistical significance. Pearson W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods
5 Enzymol. 183, 63-98.

BLAST2

Very sensitive protein sequence database searches with estimates of statistical significance. Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. Basic local alignment
10 search tool. Journal of Molecular Biology 215, 403-10.

PREDATOR

High-accuracy secondary structure prediction from single and multiple sequences. Frishman, D. and Argos, P. (1997) 75% accuracy in protein secondary structure prediction. Proteins, 27,
15 329-335. Frishman, D. and Argos, P. (1996) Incorporation of long-distance interactions in a secondary structure prediction algorithm. Prot. Eng. 9, 133-142.

STRIDE

Secondary structure assignment from atomic coordinates.
20 Frishman, D. and Argos, P. (1995) Knowledge-based secondary structure assignment. Proteins 23, 566-579.

CLUSTALW

Multiple sequence alignment. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of
25 progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22:4673-4680.

TMAP

Transmembrane region prediction from multiply aligned sequences. Persson, B. and Argos, P. (1994) Prediction of
30 transmembrane segments in proteins utilising multiple sequence alignments. J. Mol. Biol. 237, 182-192.

ALOM2

Transmembrane region prediction from single sequences.

- Klein, P., Kanehisa, M., and DeLisi, C. Prediction of protein function from sequence properties: A discriminant analysis of a database. Biochim. Biophys. Acta 787, 221-226 (1984). Version 2 by Dr. K. Nakai.

SIGNALP

- Signal peptide prediction Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering 10, 1-6.

SEG

- Detection of low complexity regions in protein sequences. Wootton, J.C., Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry 17, 149-163.

COILS

- Detection of coiled coils. Lupas, A., M. Van Dyke, and J. Stock, "Predicting Coiled Coils from Protein Sequences." Science (1991) 252, 1162-1164.

PROSEARCH

- Detection of PROSITE protein sequence patterns. Kolakowski L.F. Jr., Leunissen-J.A.M., Smith-J.E. (1992) ProSearch: fast searching of protein sequences with regular expression patterns related to protein structure and function. Biotechniques 13, 919-921.

BLIMPS

- Similarity searches against a database of ungapped blocks. J.C. Wallace and Henikoff S., (1992) PATMAT: a searching and extraction program for sequence, pattern and block queries and databases, CABIOS 8, 249-254. Written by Bill Alford.

HMMER

Hidden Markov model software . Sonnhammer E.-L.-L., Eddy S.-R., Durbin R. (1997) Pfam: A Comprehensive Database of Protein Families Based on Seed Alignments. Proteins 28, 405-420.

5 pI

Perl script that returns the amino acid composition, molecular weight, theoretical pI, and expected extinction coefficient of an amino acid sequence. By Fred Lindberg. The parameter-settings were as follows: known3d: score > 100; BLAST: E-value < 10; SCOP: 10 <= 50 Alignments, E-Value < 0.0001; signalp: Y=0.7; untersucht vom N-Terminus her: 50 aa; funcat: E-value < 0.001; BLOCKS: <= 10 hits; BLIMPS: threshold 1100.0; COILS: threshold 0.95; SEG: threshold 20.0; BLAST in report: E-value < 0.001; PIR-KW, 15 superfamilies, EC-Nummern in report: E-value < 0.00001; known3d in report: score > 120

The results of PEDANT analysis together with the results of the similarity searches constitute the basis for the structural and functional annotation of the cDNAs and the encoded proteins, as specified herein.

We claim:

1. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_12g7; amy2_12il; amy2_13g19; amy2_16el4; amy2_24kl5; amy2_2al3; amy2_2il7; fbr2_78dl8; fbr2_78el8; amy2_12lm2; amy2_24b4; amy2_12lfl9; tes3_16b5; amy2_1i24; amy2_1jl9; amy2_2bl9; amy2_7j5; amy2_14b5; amy2_2ol3; fkd2_3kl; mel2_7gl4; mel2_12jl ; mel2_7kl9; amy2_2c22; fbr2_78i2l; amy2_1ln4; amy2_1cl2; amy2_1il; amy2_2f22; amy2_2gl2; fbr2_78cl2; tes3_10il6; tes3_3la10; amy2_10hl7; amy2_10p7; amy2_12d7; amy2_2fl8; tes3_1lc22; tes3_1ld2l; tes3_29f24; tes3_3lj20; tes3_5k22; Tes3_10nl0; Tes3_1lel7; Tes3_12dl8 ; Tes3_14l7; Tes3_15nl4; Tes3_16p3; Tes3_19pl2; Tes3_2lk14; Tes3_22il1; Tes3_22l24; tes3_26g3; tes3_30pb; amy2_1ld2; amy2_12lol7; amy2_1il4; amy2_24c8; fbr2_78d4; tes3_1la17; tes3_17i2l; tes3_20hl2; tes3_7nl2; tes3_9el6; amy2_14ml6; tes3_18nl4; their complements; and variants thereof.

2. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_12g7; amy2_12il; amy2_13g19; amy2_16el4; amy2_24kl5; amy2_2al3; amy2_2il7; amy2_12lm2; amy2_24b4; amy2_12lfl9; amy2_1i24; amy2_1jl9; amy2_2bl9; amy2_7j5; amy2_14b5; amy2_2ol3; amy2_2c22; amy2_1ln4; amy2_1cl2; amy2_1il; amy2_2f22; amy2_2gl2; amy2_10hl7; amy2_10p7; amy2_12d7; amy2_2fl8; amy2_1ld2; amy2_12lol7; amy2_1il4; amy2_24c8; their complements; and variants thereof.

3. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: fbr2_78dl8; fbr2_78el8; fbr2_78i2l; fbr2_78cl2; fbr2_78d4; their complements; and variants thereof.

4. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_12lm2; amy2_24b4; their complements; and variants thereof.

5. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_121f19; tes3_1bb5; their complements; and variants thereof.

5 b. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_1i24; amy2_1j19; amy2_2b19; amy2_7j5; their complements; and variants thereof.

7. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_14b5; amy2_2ol3; fkd2_3kl; mel2_7gl4; their complements; and variants thereof.

8. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of mel2_7gl4; mel2_12jl ; mel2_7kl9; their complements; and variants thereof.

9. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_2c22; fbr2_78i21; their complements; and variants thereof.

10. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_1ln4; amy2_1il; amy2_2gl2; fbr2_78cl2; tes3_10ilb; tes3_3la10; their complements; and variants thereof.

11. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_10hl7; amy2_10p7; amy2_12d7; amy2_2fl8; tes3_1lc22; tes3_1ld21; tes3_29f24; tes3_3lj20; tes3_5k22; their complements; and variants thereof.

12. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: tes3_1bb5; tes3_10ilb; tes3_3la10; tes3_1lc22; tes3_1ld21; tes3_29f24; tes3_3lj20; tes3_5k22; Tes3_10n10; Tes3_1le17; Tes3_12d18 ; Tes3_14l7; Tes3_15n14; Tes3_1bp3;

Tes3_19p12; Tes3_21k14; Tes3_22i11; Tes3_22i24; tes3_2bg3;
tes3_30pb; tes3_11a17; tes3_17i21; tes3_20h12; tes3_7n12;
tes3_9e1b; their complements; and variants thereof.

5 13. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_11d2; amy2_12l017; amy2_1i14; amy2_24c8; fbr2_78d4; tes3_11a17; tes3_17i21; tes3_20h12; tes3_7n12; tes3_9e1b; their complements; and variants thereof.

10 14. An assemblage, comprising at least one nucleic acid molecule having the sequence of a clone selected from the group consisting of: amy2_14m1b; tes3_18n14; amy2_1c12; amy2_2f22; their complements; and variants thereof.

15 15. A nucleic acid molecule comprising a nucleotide sequence of the clone fkd2_3k1.

16. A computer readable medium, comprising in electronic form at least one nucleic acid or protein sequence of a clone selected from the group consisting of: amy2_12g7; amy2_12i1; amy2_13g19; amy2_16e14; amy2_24k15; amy2_2a13; amy2_2i17; fbr2_78d18; fbr2_78e18; amy2_12l12; amy2_24b4; amy2_12l119; tes3_16b5; amy2_1i24; amy2_1j19; amy2_2b19; amy2_7j5; amy2_14b5; amy2_2o13; fkd2_3k1; mel2_7g14; mel2_12j1; mel2_7k19; amy2_2c22; fbr2_78i21; amy2_11n4; amy2_1c12; amy2_1i1; amy2_2f22; amy2_2g12; fbr2_78c12; tes3_10i1b; tes3_31a10; amy2_10h17; amy2_10p7; amy2_12d7; amy2_2f18; tes3_11c22; tes3_11d21; tes3_29f24; tes3_31j20; tes3_5k22; Tes3_10n10; Tes3_11e17; Tes3_12d18; Tes3_14i17; Tes3_15n14; Tes3_16p3; Tes3_19p12; Tes3_21k14; Tes3_22i11; Tes3_22i24; tes3_2bg3; tes3_30pb; amy2_11d2; amy2_12l017; amy2_1i14; amy2_24c8; fbr2_78d4; tes3_11a17; tes3_17i21; tes3_20h12; tes3_7n12; tes3_9e1b; amy2_14m1b; tes3_18n14; their complements; and variants thereof.

17. A computer readable medium, comprising in electronic form at least one nucleic acid or protein sequence of a clone selected from the group consisting of: amy2_12g7; amy2_12i1; amy2_13g19; amy2_16e14; amy2_24k15; amy2_2a13; amy2_2i17;

amy2_121m2; amy2_24b4; amy2_121f19; amy2_1i24; amy2_1j19;
amy2_2b19; amy2_7j5; amy2_14b5; amy2_2o13; amy2_2c22; amy2_1ln4;
amy2_1c12; amy2_1i1; amy2_2f22; amy2_2g12; amy2_10h17; amy2_10p7;
amy2_12d7; amy2_2f18; amy2_11d2; amy2_121o17; amy2_1i14;
5 amy2_24c8; their complements; and variants thereof.

18. A computer readable medium, comprising in electronic
form at least one nucleic acid or protein sequence of a clone
selected from the group consisting of: fbr2_78d18; fbr2_78e18;
fbr2_78i21; fbr2_78c12; fbr2_78d4; their complements; and
10 variants thereof.

19. A computer readable medium, comprising in electronic
form at least one nucleic acid or protein sequence of a clone
selected from the group consisting of: amy2_121m2; amy2_24b4;
their complements; and variants thereof.

15 20. A computer readable medium, comprising in electronic
form at least one nucleic acid or protein sequence of a clone
selected from the group consisting of: amy2_121f19; tes3_1bb5;
their complements; and variants thereof.

20 21. A computer readable medium, comprising in electronic
form at least one nucleic acid or protein sequence of a clone
selected from the group consisting of: amy2_1i24; amy2_1j19;
amy2_2b19; amy2_7j5; their complements; and variants thereof.

25 22. A computer readable medium, comprising in electronic
form at least one nucleic acid or protein sequence of a clone
selected from the group consisting of: amy2_14b5; amy2_2o13;
fkd2_3k1; mel2_7g14; their complements; and variants thereof.

30 23. A computer readable medium, comprising in electronic
form at least one nucleic acid or protein sequence of a clone
selected from the group consisting of: mel2_12j1; mel2_7k19;
their complements; and variants thereof.

24. A computer readable medium, comprising in electronic
form at least one nucleic acid or protein sequence of a clone
selected from the group consisting of: amy2_2c22; fbr2_78i21;
their complements; and variants thereof.

25. A computer readable medium, comprising in electronic form at least one nucleic acid or protein sequence of a clone selected from the group consisting of: amy2_1ln4; amy2_1il; amy2_2gl2; fbr2_78cl2; tes3_10ilb; tes3_3la10; their complements; and variants thereof.

26. A computer readable medium, comprising in electronic form at least one nucleic acid or protein sequence of a clone selected from the group consisting of: amy2_10hl7; amy2_10p7; amy2_12d7; amy2_2fl8; tes3_1lc22; tes3_1ld21; tes3_29f24; tes3_3lj20; tes3_5k22; their complements; and variants thereof.

27. A computer readable medium, comprising in electronic form at least one nucleic acid or protein sequence of a clone selected from the group consisting of: tes3_1bb5; tes3_10ilb; tes3_3la10; tes3_1lc22; tes3_1ld21; tes3_29f24; tes3_3lj20; tes3_5k22; Tes3_10nl0; Tes3_1le17; Tes3_12dl8; Tes3_14l7; Tes3_15nl4; Tes3_1bp3; Tes3_19pl2; Tes3_21kl4; Tes3_22il1; Tes3_22l24; tes3_2bg3; tes3_30pb; tes3_1la17; tes3_17i21; tes3_20hl2; tes3_7nl2; tes3_9elb; their complements; and variants thereof.

28. A computer readable medium, comprising in electronic form at least one nucleic acid or protein sequence of a clone selected from the group consisting of: amy2_1ld2; amy2_12lol7; amy2_1il4; amy2_24c8; fbr2_78d4; tes3_1la17; tes3_17i21; tes3_20hl2; tes3_7nl2; tes3_9elb; their complements; and variants thereof.

29. A computer readable medium, comprising in electronic form at least one nucleic acid or protein sequence of a clone selected from the group consisting of: amy2_14mlb; tes3_18nl4; amy2_1cl2; amy2_2f22; their complements; and variants thereof.

30. A computer readable medium, comprising in electronic form a nucleic acid or protein sequence of the clone fkd2_3kl.